Predictive assessment of models for dynamic functional connectivity

Nielsen, Søren Føns Vind; Schmidt, Mikkel Nørgaard; Madsen, Kristoffer Hougaard; Mørup, Morten

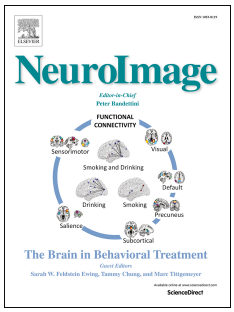
Published in:
NeuroImage

Link to article, DOI:
10.1016/j.neuroimage.2017.12.084

Publication date:
2018

Document Version
Peer reviewed version

Accepted Manuscript



Predictive assessment of models for dynamic functional connectivity

Søren F.V. Nielsen, Mikkel N. Schmidt, Kristoffer H. Madsen, Morten Mørup

PII: S1053-8119(17)31108-4

DOI: 10.1016/j.neuroimage.2017.12.084

Reference: YNIMG 14599

To appear in: *NeuroImage*

Received Date: 8 September 2017

Accepted Date: 24 December 2017

Please cite this article as: Nielsen, Sø.F.V., Schmidt, M.N., Madsen, K.H., Mørup, M., Predictive assessment of models for dynamic functional connectivity, *NeuroImage* (2018), doi: 10.1016/j.neuroimage.2017.12.084.

Predictive Assessment of Models for Dynamic Functional Connectivity

Søren F. V. Nielsen[a,1], Mikkel N. Schmidt[a], Kristoffer H. Madsen[a,b], Morten Mørup[a]

[a]*DTU Compute, Technical University of Denmark,*
Richard Petersens Plads, Building 324, DK-2800 Kgs. Lyngby, Denmark

[b]*Danish Research Centre for Magnetic Resonance, Section 714,*
Copenhagen University Hospital Hvidovre, Kettlegaard Allé 30, DK-2650 Hvidovre, Denmark

Email addresses: sfvn@dtu.dk (Søren F. V. Nielsencorr), mnsc@dtu.dk (Mikkel N. Schmidt), kristofferm@drcmr.dk (Kristoffer H. Madsen), mmor@dtu.dk (Morten Mørup)

Abstract

In neuroimaging, it has become evident that models of dynamic functional connectivity (dFC), which characterize how intrinsic brain organization changes over time, can provide a more detailed representation of brain function than traditional static analyses. Many dFC models in the literature represent functional brain networks as a meta-stable process with a discrete number of states; however, there is a lack of consensus on how to perform model selection and learn the number of states, as well as a lack of understanding of how different modeling assumptions influence the estimated state dynamics. To address these issues, we consider a predictive likelihood approach to model assessment, where models are evaluated based on their predictive performance on held-out test data. Examining several prominent models of dFC (in their probabilistic formulations) we demonstrate our framework on synthetic data, and apply it on two real-world examples: a face recognition EEG experiment and resting-state fMRI. Our results evidence that both EEG and fMRI are better characterized using dynamic modeling approaches than by their static counterparts, but we also demonstrate that one must be cautious when interpreting dFC because parameter settings and modeling assumptions, such as window lengths and emission models, can have a large impact on the estimated states and consequently on the interpretation of the brain dynamics.

Keywords: Dynamic functional connectivity, Hidden Markov models, Predictive likelihood.
[1]Corresponding author

1. Introduction

The functional integration of the brain can be studied by analyzing the patterns of synchronized activity across spatially separated brain regions. It has recently been shown that the functional connectivity (FC) varies with time, and a number of studies have investigated this dynamic functional connectivity (dFC) both in magneto/electro-encephalography (M/EEG) and functional magnetic resonance imaging (fMRI) (see recent reviews by Hutchison et al. 2013; Calhoun et al. 2014; Calhoun & Adali 2016; O'Neill et al. 2017).
dFC can be studied by computing a static measure of FC (such as Pearson correlation) for successive windowed segments of activation time series. In accordance with the idea of meta-stability in the brain, cluster analysis (e.g. using the k-means algorithm) of the dFC time courses can then be used to identify a smaller set of FC *states* that occur repeatedly across time (Allen et al., 2014). A challenge with this windowed k-means (WKM) approach is that it is sensitive to the selection of the window length (Shakil et al., 2016; Hindriks et al., 2016) which implicitly defines the time scale of the dFC.

As an alternative to WKM, a window free approach based on a hidden Markov model (HMM) has recently been proposed (Baker et al., 2014; Ryali et al., 2016; Vidaurre et al., 2017a; Nielsen et al., 2016; Vidaurre et al., 2017b). A HMM is a probabilistic sequence model which assigns a state label to each time point in the activation time series. The transitions between states are governed by a Markov process, and each state is characterized by a probability distribution over possible observations (which we refer to as the *emission model*). The state sequence, transition probabilities, and parameters of the emission model are estimated jointly when fitting the model. Analyzing resting state MEG power envelopes, Baker et al. (2014) proposed using a multivariate Gaussian emission model with state

[1]Corresponding author

specific mean and covariance. A more advanced state-specific vector auto-regressive (VAR) emission model was considered by Vidaurre et al. (2016) to model raw MEG time series, in which each state also captures frequency structure and interactions in terms of a multivariate set of autoregressive coefficients. In resting state fMRI, the HMM with Gaussian emission model has been used in Ryali et al. (2016); Vidaurre et al. (2017b). The sliding window and HMM-based approaches to modeling dFC are illustrated in Fig. 1.

Several studies have investigated the statistical support for the assumption of dynamic changes in FC. Using an auto-regressive model of pairwise connections between brain nodes, Zalesky et al. (2014) found that relatively few connections were in fact dynamic but that there was support for dFC in resting state fMRI. Using a sinusoidal model, Leonardi & Van De Ville (2015) demonstrated how spurious fluctuations in FC could arise due to model mismatch, and concluded that an appropriate window length was around 100 s based on the slowest frequency component of the BOLD signal. However, as Zalesky & Breakspear (2015) points out the sinusoidal model does not capture the correct spectral properties of fMRI data, and the conclusion is that more sophisticated generative models are needed. Laumann et al. (2016) conclude in their paper on stability of the BOLD signal that some of the dynamics observed can be attributed to head motion and subjects falling asleep in the scanner, but that some of the neural signal still remains unexplained.

While dFC analysis has become a widely accepted approach to analyze functional neuroimaging data, important open problems remain, including determining the number of brain states, and for sliding window methods to determine the window length. While a HMM can estimate the appropriate time scale from data, it is unclear how to best define the emission model. Since these modeling choices can significantly influence the interpretation of dFC, we posit there is a demand for a principled approach to compare different models of dFC.

In this paper we present a predictive model validation method in which dFC models are assessed based on their ability to characterize previously unseen data from the same experiment. To predict held-out data in a principled and quantifiable manner, we take a fully probabilistic modeling approach. While HMMs are probabilistic by nature, the WKM approach is not. We therefore consider WKM within a probabilistic setting by reformulating it as a Wishart mixture model (WMM) (Hidot & Saint-Jean, 2010; Korzen et al., 2014; Cherian et al., 2016; Nielsen et al., 2017). Within these probabilistic model specifications we use predictive validation to estimate the appropriate model complexity, including the appropriate number of brain states within each model specification. We thereby quantify whether or not functional connectivity is dynamic: ”Does the data support more than one state?”, as well as the complexity of dFC: ”How many states best account for the held-out data?” in a data-driven way. For dFC specified by HMMs, we use our predictive assessment method to systematically investigate the influence of different emission models on the number of estimated states as well as on their ability to characterize held-out functional data.

We hypothesize that dynamics in dFC-models are strongly influenced by model parameters such as window lengths, emission models, and model order, and that the more complicated emission models will be able to explain the data at hand using fewer states compared to the simpler emission models. We demonstrate this using our predictive assessment framework on both synthetic dFC data with ground truth as well as real publicly available EEG (Wakeman & Henson, 2015) and fMRI data (Poldrack et al., 2015).

Figure 1: Overview of the sliding window approach and hidden Markov model for extracting dFC. In this example both models were fitted on ERP-data from one subject (see section 3.3) using independent component analysis (ICA) time-courses of the neuronal signal from a number of brain regions. (a) In the sliding window approach, we divided the input time courses into 9 non-overlapping windows, each with 50 time points, and then computed the correlation matrix for each window. Next, we clustered the correlation matrices using k-means clustering with 3 components. (b) The hidden Markov model was fitted directly to the time courses using a multivariate Gaussian emission model and 3 states.

2. Methods

In the following, we examine four different models: a probabilistic formulation of the WKM as well as three hidden Markov models with different emission models. We treat all models in a non-parametric Bayesian setting (Orbanz & Teh, 2011): Using a prior distribution for the states based on a Dirichlet process (DP) allows us to estimate both the state parameters as well as the number of states simultaneously from data. We formulate the WKM approach as a DP-mixture model (Rasmussen, 1999) with Wishart distributed observations of windowed covariance matrices, and we consider three Gaussian DP-HMMs (Beal et al., 2002) with state-specific covariance and i) zero mean, ii) state-specific mean, and iii) state-specific vector auto-regressive mean. These non-parametric Bayesian models are commonly referred to as "infinite", as they can be derived by taking a limit as the number of states goes to infinity in a corresponding finite state model. Although these models support an unbounded number of states, inference on a finite data set will invoke only a finite subset, thus providing a statistically well founded mechanism for estimating the number of states. We further contrast this approach to the more conventional finite, parametric modeling approach as implemented by Vidaurre et al. (2016) (see also the appendix section B).

2.1. The Infinite Wishart Mixture Model (IWMM)

The windowed k-means (WKM) approach has been used extensively in the dFC literature (Allen et al., 2014; Rashid et al., 2016). Small "snapshots" of connectivity are estimated using *L* sliding windows and the snapshots are represented as correlation matrices, $\boldsymbol{\Omega}_\ell$, for each window ℓ. To find common connectivity patterns the upper triangular part of each $\boldsymbol{\Omega}_\ell$ is stacked into a vector, $\boldsymbol{\omega}_\ell$, and finally k-means clustering is performed on the collection of vectors $\{\boldsymbol{\omega}_1, \boldsymbol{\omega}_2, \ldots, \boldsymbol{\omega}_L\}$ using *K* clusters and the Euclidean distance measure. A common problem in the WKM is how to choose *K*, and heuristics such as the elbow-criterion are often used.

To be able to perform predictive validation on previously unseen data, and to learn the number of clusters as part of the model inference, we reformulate the WKM approach as a probabilistic generative model. Each windowed covariance matrix $\boldsymbol{\Omega}_\ell$ is the mean-subtracted scatter matrix, $\boldsymbol{C}_\ell$, of the data within each window segment ℓ, defined as

$$\boldsymbol{C}_\ell = \sum_t w_\ell(t)\mathbf{x}_t\mathbf{x}_t^T, \quad (1)$$

where $\mathbf{x}_t \in \mathbb{R}^p$ is the data vector at time *t* and $w_\ell(t)$ is the window function associated with the ℓth window. Under a multivariate Gaussian assumption and rectangular windows, the scatter matrices follow a Wishart distribution, and a clustering of these can be achieved using a Wishart mixture model (WMM) as proposed by Hidot & Saint-Jean (2010). We argue that the WMM is the most natural and direct probabilistic formulation of the WKM approach. We presently consider the DP-mixture version of the WMM, the so-called infinite Wishart mixture model (IWMM), as proposed by Korzen et al. (2014).

The IWMM assumes that each state has an associated covariance matrix $\boldsymbol{\Sigma}_k$, drawn from an inverse Wishart prior, and that each observed data window belongs to one of the *K* states, where *K* lies between one and the number of observations. In the DP-mixture, the prior distribution over the state assignments is given by the so-called Chinese restaurant process (CRP) (Aldous, 1985); a distribution that has support on all state assignments corresponding to all possible partitions of the observations. This yields the following generative model for the IWMM,

$$\mathbf{z} \sim \mathrm{CRP}(\alpha), \quad (2)$$

$$\boldsymbol{\Sigma}_k \sim \mathcal{W}^{-1}(\boldsymbol{\Sigma}_0, \nu_0), \quad (3)$$

$$\mathbf{C}_\ell \sim \mathcal{W}(\boldsymbol{\Sigma}_{z_\ell}, \nu_\ell), \quad (4)$$

in which $\mathbf{z}$ is the state assignment of each window, $\mathbf{\Sigma}_0$ is the prior covariance with ν_0 degrees of freedom and ν_ℓ is the degrees of freedom for the ℓ th windowed covariance matrix (in the case of a rectangular window this is equal to the window length). Due to conjugacy between the Wishart and inverse Wishart distribution we can marginalize out all the $\mathbf{\Sigma}_k$'s and carry out the inference in terms of the state assignment parameters only. In the IWMM we parameterize the prior $\mathbf{\Sigma}_0 = \eta\mathbf{I}$, in which η is a positive scaling parameter that determines the strength of the prior.

One problem still persist for WKM and IWMM, namely how to choose the window-length. We cannot compare models using predictive likelihood across different window-lengths since the likelihood function itself depends on the window length. The most flexible choice of window length is 1, in which we arrive at a likelihood function proportional to a Gaussian mixture model (GMM), but here we lose much of the stability and robustness achieved with longer window lengths. To model a slowly changing state sequence, the most natural extension is thus to use a hidden Markov model (HMM), which we discuss in the following.

2.2. The Infinite Hidden Markov Model

In neuroimaging, hidden Markov models have frequently been used for modeling dFC (Baker et al., 2014; Vidaurre et al., 2016; Ryali et al., 2016; Nielsen et al., 2016; Vidaurre et al., 2017a,b). In a manner similar to a DP-mixture model, the non-parametric version of the HMM, termed the infinite HMM (IHMM) (Beal et al., 2002), learns the number of states as part of the inference. The generative model for the IHMM can be written as,

$$b_k \sim \text{Beta}\left(1, \gamma\right), \qquad (5)$$

$$\beta_k = b_k \prod_{\ell=1}^{k-1}\left(1-b_\ell\right), \qquad (6)$$

$$\boldsymbol{\pi}^{(k)} \mid \boldsymbol{\beta} \sim \text{DP}\left(\alpha, \boldsymbol{\beta}\right), \qquad (7)$$

$$z_t \mid z_{t-1} \sim \text{Multinomial}\left(\boldsymbol{\pi}^{(z_{t-1})}\right), \qquad (8)$$

$$\theta^{(k)} \sim H, \qquad (9)$$

$$\mathbf{x}_t \sim F\left(\theta^{(z_t)}\right), \qquad (10)$$

in which γ and α are positive parameters, $\boldsymbol{\beta}$ is a vector of infinite length (in practice one needs only to work with a finite representation), $\boldsymbol{\pi}$ is the transition matrix with rows $\boldsymbol{\pi}^{(k)}$ and DP is the Dirichlet process ((18)) — a distribution over discrete probability distributions, parameterized by a base measure $\boldsymbol{\beta}$ and a concentration parameter α (for a thorough exposition of the DP, see e.g. Blei & Jordan 2006 and Van Gael 2011). The state specific parameters, $\theta^{(k)}$, are assumed to be drawn from a here unspecified prior distribution H, and data is drawn from the unspecified distribution F dependent on which state that particular data point, x_t, belongs to. A graphical model for the IHMM can be seen in Figure S.1b in the appendix.

2.2.1. Emission Models

We investigate three emission models of increasing complexity that have previously been used for modeling dFC: a zero-mean Gaussian (ZMG) (Nielsen et al., 2016), a Gaussian with a state-specific mean (SSM) (Rezek & Roberts, 2005; Baker et al., 2014), and Gaussian with an auto-regressive mean (VAR) (Fox et al., 2011; Vidaurre et al., 2016). In all cases the covariance is state-specific and models that state's functional connectivity. There are other emission models in the Gaussian family such as the state specific mean model with isotropic variance (Baldassano et al., 2017) and other variants where the covariance is constrained. These will not be considered presently because they do not

model the full functional connectivity. The emission parameters are distributed as described in Table 1.

Zero Mean Gaussian	State-Specific Mean	Vector Autoregressive
ZMG	**SSM**	**VAR**
$\boldsymbol{\Sigma}^{(k)} \sim \mathcal{W}^{-1}\left(\boldsymbol{\Sigma}_0, \nu_0\right)$	$\boldsymbol{\Sigma}^{(k)} \sim \mathcal{W}^{-1}\left(\boldsymbol{\Sigma}_0, \nu_0\right)$	$\boldsymbol{\Sigma}^{(k)} \sim \mathcal{W}^{-1}\left(\boldsymbol{\Sigma}_0, \nu_0\right)$
	$\boldsymbol{\mu}^{(k)} \sim \mathcal{N}\left(\boldsymbol{\mu}_0, \lambda^{-1}\boldsymbol{\Sigma}^{(k)}\right)$	$\mathbf{A}^{(k)} \sim \mathcal{MN}\left(0, \boldsymbol{\Sigma}^{(k)}, \mathbf{I}\right)$
$\mathbf{x}_t \sim \mathcal{N}\left(0, \boldsymbol{\Sigma}^{(z_t)}\right)$	$\mathbf{x}_t \sim \mathcal{N}\left(\boldsymbol{\mu}^{(z_t)}, \boldsymbol{\Sigma}^{(z_t)}\right)$	$\mathbf{x}_t \sim \mathcal{N}\left(\mathbf{A}^{(z_t)}\overline{\mathbf{x}}_t, \boldsymbol{\Sigma}^{(z_t)}\right)$

Table 1: Emission models used in the HMMs where $\boldsymbol{\Sigma}^{(k)}$ is the state-specific $p \times p$ covariance matrix, $\mathcal{W}^{-1}$ is the inverse Wishart distribution, Σ_0 is the prior covariance, ν_0 is the degrees of freedom (in all experiments $\nu_0 = p$), $\boldsymbol{\mu}_0$ is the prior mean of the signal, λ is a positive precision parameter of the mean, $\mathcal{MN}\left(M, U, V\right)$ is the matrix-normal distribution with mean *M*, row-variance *U* and column variance *V*, $\mathbf{A}^{(k)}$ is a $p \times pr$ matrix containing the coefficients for the k'th state of an order *r* VAR process, and $\overline{\mathbf{x}}_t$ are the *r*-lagged observations for time point *t* stacked in a vector.

For all the HMM emission models we have chosen conjugate distributions, to be able to analytically marginalize $\boldsymbol{\Sigma}^{(k)}$, $\boldsymbol{\mu}^{(z_t)}$, and $\mathbf{A}^{(z_t)}$, such that inference is carried out on the state sequence alone.

2.3. Predictive Likelihood

To assess and compare the different models, we examine their ability to generalize, i.e., how well a model fitted on training data, $\mathbf{X}$, can account for unseen test data, $\mathbf{X}^*$, from the same experiment or paradigm. This can be viewed as an alternative to classical statistical inference and hypothesis testing (Bzdok & Yeo, 2017).

Thus we are interested in evaluating the following integral,

$$p\left(\mathbf{X}^* \mid \mathbf{X}, \mathcal{M}\right) = \int_{\boldsymbol{\Theta} \in \mathcal{M}} p\left(\mathbf{X}^* \mid \boldsymbol{\Theta}\right) p\left(\boldsymbol{\Theta} \mid \mathbf{X}\right), \qquad (11)$$

yielding the *posterior predictive likelihood* (from now on denoted the predictive likelihood), in which $\boldsymbol{\Theta} \in \mathcal{M}$ is the collection of all model parameters and $\mathcal{M}$ is the model space. The predictive likelihood quantifies the probability of observing the test data under the model given the training data and the model space, and can be viewed as the likelihood of the test data averaged over the posterior distribution of the parameters fitted on the training data. To showcase that the predictive likelihood framework is also applicable for other probabilistic models we also use the (finite) variational Bayesian HMM (VB-HMM) from Vidaurre et al. (2016)[2]. A description of the model can be seen in the appendix section B, along with details on how to calculate the predictive likelihood for all models.

To use predictive evaluation, the data must be divided into independent training and test sets. In dFC, where the data is modeled as sequential, this can be done by splitting the time series into sub-sequences. Alternatively, it is possible to train the model on whole time series from one or more subjects, and use data from independent, held-out subjects for testing. In this paper we use the predictive likelihood to do model selection and parameter tuning in a two level cross-validation framework. In the inner part, we estimate the prior strength η for the IWMM and IHMMs considered, and the number of states for VB-HMM for all emission models. In the outer part, we quantify each of the emission model's capability of explaining the held-out test data. We emphasize that we cannot directly compare the predictive likelihood across IWMM, VB-HMM and IHMM. The IWMM uses a

[1]different likelihood function than the two HMM-models, i.e. the IWMM models covariance matrices as the observed quantity whereas the HMMs model the time series directly. For the VB-HMM we have chosen to use a VB-bound to approximate the integral in (11) (Beal, 2003), that has the advantage of propagating the uncertainty in the parameters from training at the cost of estimating the state-sequence distribution on the test set. In the IHMM we use samples from the posterior obtained during training together with Viterbi integration (more details on this can be found in Appendices B, C and D). A general schematic of the predictive likelihood framework can be seen in Figure 2.

Of present interest is to investigate under a given independent component analysis (ICA) representation which model of dFC most adequately describes this representation. We therefore consider the ICA as a preprocessing step applied to all the data. Alternatively, the ICA could have been applied separately on the training and test data. Training the ICA independently on the training and test set would result in an issue of matching components (Du et al., 2012), whereas defining the ICA only on the training data and projecting the test data onto these learned components can result in issues of variance inflation (Abrahamsen & Hansen, 2011). By considering the ICA as a preprocessing step we remove any influence that changes in the ICA representation across training and test data may have. We are thereby not affected by these potential confounds and are able to quantify within a given ICA representation which model of dFC best characterizes the data.

For the remainder of this paper we will contrast the predictive likelihood of a model of interest versus a baseline model using the Bayes factor (Kass & Raftery, 1995; Nielsen et al., 2017), denoted BF. This can be calculated as,

$$\mathrm{BF} = \frac{p\left(\mathbf{X}^* \mid \mathbf{X}, \mathcal{M}\right)}{p\left(\mathbf{X}^* \mid \mathbf{X}, \mathcal{M}_0\right)}, \quad (12)$$

in which $\mathcal{M}$ is the model of interest and $\mathcal{M}_0$ is the baseline model. Typically the baseline model will be the model with only one state, and thus the Bayes factor quantifies how much better it is to use a particular dynamic model. The Bayes factor is often used in the dynamic causal modeling (DCM) framework (Penny et al., 2004) to do model selection, however, an important distinction between the DCM and our approach is that the BF in DCM in calculated on the training data whereas the BF in this paper is calculated on held-out test data.

Figure 2: A schematic overview of the predictive likelihood framework, that visualizes the nested cross-validation framework. In this figure the models of dynamic functional connectivity (dFC) can be anything, as long as the predictive likelihood on held-out data can be estimated. If the likelihood function is the same across models we can use this framework to do model selection.

2.4. Evaluating similarity of state sequences

To compare different models, we also examine how similar their estimated state sequences are. Here, we use normalized mutual information (NMI) to quantify the correspondence of two different sequences, possibly with differing number of states. Let the state sequences be given by state assignments vectors $\boldsymbol{z}^{(a)}$ and $\boldsymbol{z}^{(b)}$. Then, the NMI is given by

$$\mathrm{NMI}\left(\boldsymbol{z}^{(a)}, \boldsymbol{z}^{(b)}\right) = \frac{2\mathrm{MI}\left(\boldsymbol{z}^{(a)}, \boldsymbol{z}^{(b)}\right)}{\mathrm{MI}\left(\boldsymbol{z}^{(a)}, \boldsymbol{z}^{(a)}\right) + \mathrm{MI}\left(\boldsymbol{z}^{(b)}, \boldsymbol{z}^{(b)}\right)}, \quad (13)$$

where MI is the mutual information.

[2]MATLAB code was downloaded from the repository https://github.com/OHBA-analysis/HMM-MAR in July 2016. The predictive likelihood code was written by the authors.

[3]The data were downloaded from ftp://ftp.mrc-cbu.cam.ac.uk/personal/rik.henson/wakemandg_hensonrn/ including preprocessing scripts for SPM.

[4] http://www.fil.ion.ucl.ac.uk/spm/software/spm8/

3. Experiments and results

The proposed approach for predictive assessment of dFC models was validated on synthetic data, and demonstrated on two real data sets based on electroencephalography (EEG) and functional magnetic resonance imaging (fMRI) as described in the following sections.

3.1. The influence of window lengths

A challenge in the WKM approach as well as its probabilistic formulation, the IWMM, is the specification of a suitable window length (Shakil et al., 2016; Zalesky & Breakspear, 2015; Leonardi & Van De Ville, 2015; Hindriks et al., 2016). If the window length is too short, the windowed data will be less statistically stable and the approach might find spurious states. If, on the other hand, the window length is too long, short-lived states might not be detectable. In contrast, the HMM approach does not involve windowed analysis.
We applied WKM as well as the IWMM and IHMM to synthetic data with ground truth in order to investigate the merits of windowed covariance modeling versus HMMs that do not assume a priori time-windowing but learns the state dynamics and their smoothness as part of the inference. In all analysis the AR-order was set to 1 in both the IHMM and the VB-HMM (see more about this choice in the Discussion).

Synthetic data I.

We generated two data sets (training and validation) from the same 5-dimensional 3-state ZMG model, i.e., the data were generated to have three different states, defined by different covariance matrices, in a fixed state sequence. The covariance matrices for each state were generated as $\mathbf{U}\mathbf{U}^\top$ where $\mathbf{U}$ was an upper triangular matrix with i.i.d. standard Gaussian entries. In each data set, the total length of the generated time series was 500 samples (i.e., if this was fMRI we would have had 500 TRs), and the state sequence was chosen such that the states had different durations with the shortest state occurrence lasting 50 samples.

Synthetic experiment I.

For WKM we set the number of states to the true number of states ($K = 3$). For IWMM and IHMM we tuned the prior covariance scale parameter η by fitting the models on the training data and optimizing the parameter using predictive likelihood on the validation data. We then concatenated the training and validation data, and using the full data set with 1000 time points we fitted the WKM, IWMM, and IHMM-ZMG models. For the WKM and the IWMM we used rectangular non-overlapping windows, and compared window lengths of 5, 25, and 100 samples chosen to represent a too short, an appropriate (i.e. one that does not mix together different states), and a too long window.

Results on synthetic data I.

The results can be seen in Figure 3 which shows the estimated state sequences. The WKM and IWMM perform almost identically: When the window length is appropriate (WL=25) both methods detect the correct state sequence. When the the window length is too large (WL=100) both fail to capture the short-lived state correctly, and when it is too small (WL=5) the WKM detects spurious states. Both IWMM and the IHMM-ZMG correctly identify the number of states using the cross-validated value of η. Furthermore, the IHMM captures the true state sequence without a priori specifying and averaging across windows. Thus, all models can correctly identify the underlying dFC on data in compliance with their assumptions. It should be noted that adequately tuned *overlapping* and *tapered windows* (Allen et al., 2014) could potentially alleviate the issues encountered using too long window lengths, however, this was not considered in this experiment.

Figure 3: Estimated and true state sequences for synthetic data I experiment. Data were generated from a three-state model, where each states had a differing covariance matrix. Results are shown for windowed k-means (WKM) and infinite Wishart mixture model (IWMM) with window lengths of 5,

25, and 100 samples as well as the infinite hidden Markov model with zero mean Gaussian emission model (IHMM-ZMG).[2]

3.2. HMM emission models

In the hidden Markov model approach to estimating dFC, we claim that the choice of emission model can have a large influence on the result. To substantiate this, we compared the three examined emission models by performing a pair-wise comparison investigating how well each model was able to estimate the true state sequence on synthetic data generated according to each of the three model specifications. Furthermore, we compared how well each model was able to characterize dFC by computing the predictive likelihood on held-out validation data.

Synthetic data II.

We generated synthetic data from each of the three emission models (ZMG, SSM, and VAR) with five dimensions and three states (we used the same state sequence as in the previous synthetic experiment shown in Figure 3). Training, validation, and test data sets were generated with identical parameter settings for each data model. For all models, the covariance matrix for each state was defined as in the previous synthetic experiment. For the SSM model, the state-specific means (5-dimensional vectors) were generated randomly with i.i.d. standard Gaussian entries. The state-specific VAR coefficients were generated, by first generating a p-dimensional signal from a sinusoid with random frequency (common for all dimensions) and random phase (different for each dimension). We then fitted a VAR-model of order 1 to that (using the least squares estimator) and finally generated new data from the fitted model with i.i.d. standard Gaussian noise.

Synthetic data experiment II.

For IWMM and IHMM, the prior strength η was selected by cross-validation using the training and validation set, and the models were then fitted on the concatenated training and validation data. The predictive likelihood was computed for each of the fitted models using the test data. For comparison we also fitted the WKM model, both with the correct number of clusters ($K = 3$) and with too many clusters ($K = 6$). Both the WKM's and IWMM were run with an appropriate window length of 25. To investigate the influence of the inference procedure, we also fitted the models using the VB-HMM implementation by Vidaurre et al. (2016).

Results on synthetic data II.

The estimated state sequences for each of the fitted models are shown in Figure 4. When the number of states was specified correctly ($K = 3$) the WKM found the true state sequence for all three data sets; however, when the number of states was mis-specified ($K = 6$) the WKM failed in all cases and appeared to subdivide each state. The IWMM was able to learn the true state sequence for the ZMG and SSM-emission data, but failed in the case of the VAR-emission data. The three IHMM models found the true state sequence in the cases when the data were generated from one of the two simple emission models (ZMG and SSM), except the IHMM-VAR which falsely detected two single-time-point clusters for the SSM-data. When the data were generated from the VAR model, only the VAR model and the WKM with the correct number of clusters found the correct state sequence. In this setting, the IHMM-ZMG and IHMM-SSM both failed in estimating the true number of underlying states and detected multiple spurious states. This indicates that these more simple models needed more states (and parameters) to account for the more complex VAR data. Results for VB-HMM were similar to the IHMM and can be found in the appendix section E.

The predictive likelihood of each model is reported in Figure 5, which shows the predictive Bayes Factor of each emission model vs. a baseline model given by a one state (non-dynamic) zero mean Gaussian defined by the empirical covariance matrix of the concatenated training and validation set. As expected when the HMM emission model matched the emission model of the generated data, the

[5]http://www.fieldtriptoolbox.org/
[6]https://openfmri.org/dataset/ds000031/
[7]http://www.fil.ion.ucl.ac.uk/spm/software/spm12/

best Bayes factor was achieved. When the data were simple (from ZMG) the three emission-models performed approximately equal (with the ZMG performing best), indicating that the more complex models could adapt to the simple data but not vice versa. We also conducted an experiment to investigate the influence of noise and fMRI signal properties on the predictive results. This can be seen in the appendix section G.

Figure 4: Estimated state sequences for synthetic data II generated from hidden Markov models. Top: Zero mean Gaussian (ZMG) emission. Middle: State-specific mean (SSM) emissions. Bottom: Vector autoregressive (VAR) emission. Results are shown for windowed k-means (WKM) with $K = 3$ and $K = 6$ clusters, the infinite Wishart mixture model (IWMM), and infinite hidden Markov models with ZMG, SSM, and VAR emission models. The true state sequence is shown in Figure 3.

Figure 5: Bayes factors (synthetic data) for each model vs. a baseline model containing only one zero-mean state (static functional connectivity) computed on held-out validation data. This was one using both Markov chain Monte Carlo inference (left) and variational Bayesian inference (right).

3.3. EEG task paradigm analysis

To verify that the proposed predictive evaluation framework produces sensible results, we demonstrate it on an electroencephalography (EEG) task-paradigm with very high signal-to-noise ratio using event related potentials (ERP), similar to the analysis carried out in Murray et al. (2008); Ott et al. (2011) under the name of topographical ERP mapping.

EEG data.

We analyzed a publicly available face recognition task data set (Wakeman & Henson, 2015)[3] that consists of 16 subjects. The paradigm has three conditions: Either i) a *famous* face is presented, ii) an *unfamiliar* face is presented, or iii) a *scrambled* face (with the phase of the 2D-Fourier coefficients permuted) is presented. Our analysis was not focused on contrasting the conditions, and each condition was thus analyzed individually to investigate the robustness of the estimated dynamics.

The standard preprocessing, as described by Wakeman & Henson (2015) (which included low-pass filtering to 32 Hz), using the SPM8 MATLAB toolbox[4] was applied to the data and additionally we interpolated the automatically detected bad channels using the distance function in FieldTrip [5]. We then calculated individual event-related potentials (ERP) for each subject and condition and ran independent component analysis (ICA) on the concatenated data (all subjects and conditions) using the Infomax algorithm (Bell & Sejnowski, 1995) with five components. The number of components was chosen based on the eigenvalue spectrum of random uncorrelated data (Horn, 1965). An example of an ICA time course is displayed in lower right corner of Figure 7.

EEG experiment.

Eleven subjects were taken out for training, leaving five subjects for testing. In the training set five-fold cross-validation was applied to estimate the prior strength in the IHMM and the number of states for VB-HMM for each condition and each emission model using predictive log-likelihood on the validation set as a measure of fit. Each subject's ICA time courses from event related potentials (ERP) were concatenated in time, and to account for discontinuities in the data we set up the models to restart the state sequence at each new subject. After cross-validation, we re-trained the models on the whole training data and calculated the predictive likelihood on the test data.

To assess the robustness of the approach, we computed the normalized mutual information (NMI) of the estimated state sequences over five restarts of each model in the following manner: Restart (1 vs. 2), (2 vs. 3), (3 vs. 4), (4 vs. 5), and (5 vs. 1). To examine the similarity between the estimated state sequences across the three models, we computed the NMI between the models: Restart (1 vs. 1), (2 vs. 2) etc. for each pair of models (ZMG vs. SSM), (ZMG vs. VAR), and (SSM vs. VAR). As a baseline, each case was also compared to a null-model, in which one of the state sequences in each pair was

[3]replaced with a new state sequence sampled using the fitted transition matrix thus resulting in similar state transition dynamics as the original sequence but uninformed by the data.

EEG results.

NMI scores comparing the estimated state sequences are given in Figure 6a. Results for the three data sets (*familiar*, *unfamiliar*, and *scrambled*) were generally in close agreement with each other. For all models, NMI scores between restarts were higher than the baseline; thus, the estimated state sequences were relatively consistent over restarts, although all NMI scores were well below one, indicating some disagreement. NMI scores between ZMG and SSM were similar to NMI scores between restarts of the two models, indicating that the ZMG and SSM models estimated similar state sequences. NMI scores between VAR and the other two models were lower than NMI between restarts, indicating that the estimated state sequences for the VAR model were different from those estimated by the ZMG and SSM models. This was confirmed for the VB-HMM when running the ZMG and SSM models with the same number of states as the VAR model was run with, i.e., the state sequence obtained from the VAR model differs from the ZMG and SSM state sequences even if the ZMG and SSM have the same number of states as the VAR model. We also looked into the number of states estimated by each emission model; the VAR model estimated fewer states as hypothesized in the introduction with more smooth trajectories compared to the two other emission models (see the appendix section J).

Figure 6: Normalized mutual information (NMI) between estimated state sequences (circles) for each data set and each pair of models. Within each model, the NMI measures the consistency of the estimated state sequences across five reruns of the inference algorithm. Between each pair of models, the NMI measures the similarity between the estimated state sequences. NMI computed against a random state sequence from the fitted model is shown as a baseline (crosses). Results are shown for inference using MCMC (left) and variational Bayes (right). For the variational Bayes (VB) models, the VAR was also compared to the ZMG and SSM model run with the same number of states as the VAR indicated by plusses in the third row of the VB-plot.

Figure 6: Comparison of model performance on the EEG-data. We plot the IHMM and VB-HMM performance in terms of model consistency and predictive likelihood on held-out data.

To investigate which emission model best characterized the held out subjects, the Bayes factor towards a baseline model (empirical covariance matrix of the training data) was calculated and can be seen in Figure 6b. All models gave better performance than the baseline. The VAR emission model consistently gave best predictive performance across all conditions and for both inference methods.

For the best performing model, the IHMM-VAR, we take the best sample in terms of joint log-likelihood from the training inference, and visualize the solution in Figure 7. To plot the topography of each state, we gathered all the time points assigned to a particular state and calculated the first principal direction, and plotted those values using EEGLAB. We did this to not be influenced by changes in polarity and because it resembles the microstate-analysis done in Khanna et al. (2015). We notice that there seems to be a baseline state (state 1), that some of the training subjects visit before stimulus and around 0.4 seconds after stimulus. In the period after stimulus (from 0.1 - 0.4 seconds) the dominant states' topography show high activity in the posterior areas consistent with the visual task. There seems to be a "consensus" of fewer states in the baseline (pre-stimulus and after 0.4 seconds after stimulus) and a larger number of different states being used right after stimulus. This indicates that we need more states to explain the difference in visual processing of faces across subjects compared to the baseline state. Furthermore, some states seem to have very similar topographical characteristics (i.e. states 3-5) but are different in their functional connectivity.

Figure 7: Visualization of the best IHMM model, in this case the IHMM-VAR, according to the predictive framework for the "Famous" condition. (a) For each state we computed the first principal component of all the data points in the training set that belonged to that state and plotted that as a

[8]http://mialab.mrn.org/software/gift/index.html
[9]https://brainconnectivity.compute.dtu.dk/

topographical map. Note that the states were ordered according to their fractional occupancy (largest state first). Below each map we plot the empirical correlation matrix of all data points assigned to the given state. (b) We plot for each timepoint the state-assignment for each subject as an image (each row represents a subject). Each color represents a state. (c) An example of one subjects data in ICA-space (each color represents a independent component).

3.4. fMRI resting state analysis

Finally, we will demonstrate our approach to predictive assessment of dFC models on a resting state fMRI data set. Subject variability can be a significant issue in dFC (Nielsen et al., 2016) and in neuroimaging in general (Finn et al., 2015) and care must be taken when interpreting dynamics at a group level, so we analyzed resting state fMRI data from a single subject. We contrast the extracted brain states from the HMM framework to those from sliding window k-means.

fMRI data.

We used the resting state fMRI data from Poldrack et al. (2015)[6] which contains 89 recorded resting state fMRI sessions of a single subject. We applied the following pre-processing steps using SPM12[7]: We coregistered all sessions to the first image of the first functional session (session 014), and then jointly corrected all sessions for motion artifacts using a rigid-body transformation towards the mean volume. An anatomical image (T1W) from session 012 was used to segment grey matter (GM), white matter (WM) and cerebrospinal fluid (CSF) using the standard tissue probability map from SPM. We applied a discrete cosine transform based bandpass filter with cut-off at $[0.009, 0.08]$ Hz to all sessions (as suggested in the methods section of (Poldrack et al., 2015)), along with nuisance regression of the motion parameters and mean signal within CSF and WM masks eroded by a 4mm isotropic spherical kernel. We subsequently applied wavelet despiking (Patel et al., 2014) with standard parameters, and finally we resliced all sessions (due to a change in the number of slices after session 027) to the first session and smoothed using an isotropic 5mm full width at half maximum Gaussian kernel. After preprocessing we ran a group ICA ((13)) implemented in the GIFT toolbox[8], using the ERBM algorithm with 30 components and otherwise default settings. We used 30 components, which can seem 'low' compared to other dFC analysees (Allen et al., 2014). However, this was done both for computational reasons, i.e. the HMM scales cubicly in the number components (cf. appendix B) and also for statistical reasons since we need enough degrees of freedom to reliably estimate the covariance matrix of each state. We discarded 9 components based on visual inspection of the component spatial maps overlap with the brainstem and movement related effects, and thus ran the final HMM-analysis on the 21 remaining components. The retained components' spatial maps can be seen in the appendix section K.

fMRI experiment

The data were only analyzed using the VB-HMM inference procedure due to the higher computational complexity of the IHMM. We split the 89 sessions randomly into two parts: 45 sessions for training and 44 sessions for testing. In the training set we performed five fold cross-validation to determine the number of states for all three emission models using the proposed predictive log-likelihood as a measure of fit. The final models were retrained five times on the training data, the best restart chosen by the minimum free-energy, and finally compared with predictive log-likelihood on the test sessions. To compare the estimated state sequences and assess the robustness of the approach, we conducted a NMI analysis as in the EEG experiment.

fMRI results

NMI scores comparing the estimated state sequences are shown in Figure 8a. The NMI between state sequences estimated by the ZMG and SSM models were similar to NMI scores for restarts of the two models, indicating that the estimated state sequences were in agreement. NMI scores between VAR and the other two HMMs were lower, indicating that the VAR model found a different state sequence. We looked into the number of states estimated by the three emission models (see appendix section I) and found that the VAR identified six states, whereas the ZMG and SSM used 7 and 8 states respectively. From Figure 8a it seems that the WKM found more robust results over restarts and was in very low agreement with the HMMs.

Figure 8: Comparison of model performance on the fMRI-data. We plot the VB-HMM performance in terms of model consistency and predictive likelihood on held-out data. For the model consistency in 8a we compare the HMMs to windowed k-means (WKM).

The predictive performance on the test set for each of the models is given in Figure 8b, which shows log Bayes factors against a baseline given by the ZMG one state model. As in the EEG analysis the VAR model outperformed the other models in terms of predictive likelihood.

Finally, we visualize the states from best performing model (i.e. the VB-HMM VAR), by computing the mean activity of all the timepoints assigned to the same state. This is shown in Figure 9 together with the FC-matrix pr. state, a bar plot of the fractional occupancy and mean lifetime (cf. appendix H) of each state. The states' spatial activity seems to resemble the default mode network (state 3 in particular) and the sensory motor network (state 2 and 4). We note that the states seem to have a mean lifetime in the range of 10-20 TR's (10-25 sec) and that the transition matrix has a very diagonal structure indicating a lot of self-transitions, i.e. it is more likely to stay in the same state than jump to another state. Looking at the FC matrices all states seem to have a very diagonal structure with low variability over states.

Figure 9: Final VB-HMM VAR solution initialized with 7 states (one was emptied during training). The mean activity of each HMM-state is plotted, i.e. the mean of all time-points assigned to the same state. Furthermore, the empirical $p \times p$ correlation matrix for each state is also plotted (after Fisher transformation), where p is the number of ICs used (see appendix K). The states were sorted according to their fractional occupancy. Cut-coordinates were determined using Nilearn by finding the largest positive connected component after thresholding at the 95th percentile of the absolute values in the map. The fractional occupancy, mean lifetime and the transition probabilities between states is furthermore in the rightmost column.

For comparison, we ran the sliding window k-means approach (WKM) on the fMRI data with the same number of states as the VB-HMM-VAR estimated in the final run (i.e. 6 states). We used a tapered window with a window length of 22 TRs (corresponding to around 25s) sliding the window one TR at a time. We used the default MATLAB k++ initialization procedure (Arthur & Vassilvitskii, 2007) with Euclidean distance measure, and restarted the k-means procedure 100 times. However, we did not use ℓ_1-regularization as suggested in original WKM article (Allen et al., 2014), due to the well-posedness of the correlation matrices induced by the fairly low dimensionality of the problem.

For the WKM, the six states' mean activity, FC matrix and state characteristics are plotted in Figure 10. The DMN activity seems to be separated over all the states. Looking at the mean lifetime of the states we see a very uniform distribution around 20 TRs, i.e. all states seems to have the same mean lifetime, which is probably mainly due to the window length. The FC notably varies more over states compared to the VB-HMM-VAR solution in Figure 9.

Figure 10: Final WKM solution initialized with 6 states. The mean activity of each WKM-state, i.e. the mean of all time-points assigned to the same state, is plotted. Furthermore, the empirical $p \times p$ correlation matrix for each state is also plotted (after Fisher transformation), where p is the number of ICs used (see appendix K). The states were sorted according to their fractional occupancy prior to visualisation. Cut-coordinates were determined using Nilearn by finding the largest positive connected component after thresholding at the 95th percentile of the absolute values in the map. The fractional occupancy, mean lifetime and the transition probabilities between states is furthermore in the rightmost column.

4. Discussion

We have proposed a data-driven predictive framework for comparing and measuring generalization of dynamic functional connectivity (dFC) models. Using this framework we investigated a windowed covariance approach based on the infinite Wishart mixture model (IWMM) as well as the (window free) infinite HMMs (IHMM) specified by three different emission models (Nielsen et al., 2016; Baker et al., 2014; Vidaurre et al., 2016). We find that the extracted dynamics are heavily influenced by modeling assumptions. In synthetic data, where ground truth state sequences were available, it was

[4]clear that a misspecification of the model leads to an incorrect state sequence. Thus, we need to properly quantify how well certain model assumptions comply with the data observed. Here, the predictive assessment framework is able to quantify the number of states and appropriate emission model. We found the WKM to be robust towards model mismatch, however, we here in general have no a priori knowledge of either window length or the number of states that need to be specified. We found that the IWMM admits quantification of number of states within a WKM type of framework, but the choice of window length remains unresolved and influences results as illustrated in the synthetic study.

Hidden Markov models (HMMs) seem like a promising framework to circumvent the need to specify window lengths, learning state transitions and their smoothness as part of the inference. We considered both MCMC and variational Bayesian inference and consistently found that the choice of emission model heavily influences the identified functional dynamics and their interpretation as different emission models drive different dynamics. Our predictive framework admits quantification of the type of emission model that is most adequate for the system under consideration and our results points towards the vector autoregressive (VAR) model being a more flexible and better overall choice. It should be noted that in analysis of real data (EEG and fMRI) the data sets were lowpass and bandpass filtered respectively as part of the preprocessing, which may harm the estimated dynamics by driving the VAR-states towards characterizing properties of the preprocessing. In slowly fluctuating signals a large portion of the signal at time t can be explained by the signal at time $t-1$ which is exactly what the VAR(1)-model is doing in contrast to the other emission models (see also appendix section G). Preprocessing influences the estimated dynamics as shown in Hindriks et al. (2016). In this work we chose the default preprocessing pipelines as suggested by Wakeman & Henson (2015) and Poldrack et al. (2015), however we expect different preprocessing choices can favor different emission models. However, investigating these choices are out of scope of the current study. In our analyses of the EEG data (as well as in our synthetic study) there was a clear indication that the simpler HMMs (ZMG and SSM) overestimated the number of states, whereas in the analyses of the fMRI data all emission models were more in agreement. We attribute this difference to the differences in signal-to-noise ratios and temporal resolutions, but note that this requires further investigation.

On the fMRI data we compared the HMM-VAR with the WKM visualizing the brain states extracted by the two frameworks (with the same number of states). It is clear that they find different brain state representations both in mean activity, FC and temporal characteristics. As such, the WKM finds more distinct states in terms of FC than the HMM-VAR. We attribute this to their different modelling assumptions, i.e. VB-HMM VAR is a model that generates data at the level of single time points whereas the WKM is driven by characterizing differences in the off-diagonal elements of the windowed covariance matrices. When looking at the lifetimes of the extracted states, all WKM states had approximately the same length dictated by the window length used, whereas the HMM-VAR due to its window-free approach estimated states with varying lifetime. This exemplifies that dynamics are driven by the underlying model assumptions. One could be tempted to interpret what the extracted states' represent in terms of brain function, however, the NMI results in Figure 8a points toward issues with local minima in particular for the HMMs. We speculate that current dFC approaches are too flexible hampering the reliability (Choe et al., 2017), thus there seems to be a need for better inference procedures and constrained models promoting both reliability and generalization.

We compared two inference methods for the HMMs, namely Markov chain Monte Carlo (MCMC) in the form of the infinite hidden Markov model (IHMM) and variational Bayes hidden Markov model (VB-HMM). From a theoretical point of view the IHMM has the most desirable properties, i.e., we do not need to specify the number of states and we should obtain better estimates of the posterior distribution. However, in practice the IHMM and VB-HMM yield similar results, and if we factor in the computational complexity of the IHMM, the VB-HMM seems like the better choice in most practical applications.

[10]http://nilearn.github.io/

Our results supports the conclusion that functional connectivity is best modeled using multiple states (Hutchison et al., 2013; Calhoun et al., 2014; Calhoun & Adali, 2016; Vidaurre et al., 2017b). In particular, our predictive assessment consistently finds support for functional neuroimaging data, i.e., fMRI and EEG data, are better accounted for by dynamic models (i.e., models having more than one state) which was consistently observed across models and data sets. As hypothesized we find that the more advanced HMM-VAR extracted fewer states than the simpler ZMG and SSM emission models. Thus, in theory a very complicated emission model (that we have not investigated here) could potentially capture everything as "one state".
There has recently been a lot of focus on null-models and stationarity in dFC (Zalesky & Breakspear, 2015; Laumann et al., 2016; Miller et al., 2017). For choosing an appropriate window-length in WKM the work of Zalesky & Breakspear (2015) provides some statistical analysis as to why the rule of thumb of 100 second windows from (Leonardi & Van De Ville, 2015) makes sense. Zalesky & Breakspear (2015) furthermore points out that the framework can detect changes in FC on shorter timescales (around 40 seconds); changes that can disappear if longer windows are used. Their conclusion also being that we need better generative null-models for dFC. While we do not claim that we have found the true null-model for dFC, we have demonstrated a framework that admits a comparison between models based on predictive likelihood. We compared the WKM with HMM-framework on fMRI data in qualitative way; however, since the WKM is not a model of data we cannot in an objective way compare the performance of the two models. Bzdok & Yeo (2017) argues that neuroscience is moving more and more towards out-of-sample generalization as an alternative to classical statistical inference and hypothesis testing, and we will argue that models of dFC will be more objectively comparable if they are generative and are able to extrapolate to held-out data. Importantly, the HMM is a generative model that contains the static model as a special case and by doing model order selection we test in a data-driven way whether or not the FC should be modeled static ($K = 1$) or dynamic ($K > 1$).

A very important point is that the proposed framework will only answer what model best explains the data at hand. To truly validate that the extracted dynamics correspond to neurophysiological mechanisms, we need more elaborate validation such as concurrent EEG-fMRI data or even invasive studies.

Our predictive assessment framework generalizes to arbitrary dynamic model specifications as long as a predictive likelihood can be calculated. For instance, the AR-order was fixed to one in this paper but could easily be learned using the framework presented (cf. appendix F). In this paper we also show two ways of using the predictive assessment framework promoting two different kinds of generalization, i.e. between-subject generalization and within-subject generalization. We are not claiming in any way that one should use one over the other, only that we have the power with this framework to investigate both types of generalization. The quantitative analysis of this paper points to dFC being heavily influenced by modeling assumptions and the proposed assessment provides a principled tool for future refinement and tailoring of models of dFC to better account for functional neuroimaging data.

Acknowledgements

Søren F.V. Nielsen, Mikkel N. Schmidt and Morten Mørup were supported by Lundbeckfonden (fellowship grant R105-9813 to Morten Mørup). Kristoffer H. Madsen was supported by a Novo Nordisk Foundation Interdisciplinary Synergy Grant (NNF14OC0011413)

Appendices

A. Implementation details for IHMM

Both the IWMM and IHMMs were implemented using collapsed Gibbs sampling with split-merge proposals (Jain & Neal, 2004). α in the IWMM was inferred using random walk MCMC. The IHMMs were implemented on top of the the MATLAB implementation made by Van Gael (2010), in which α and γ were sampled by placing vague Gamma priors on them. As pointed out in the literature (Van Gael et al., 2008) the Gibbs sampler has some mixing issues, so to overcome this we implemented a

split-merge sampling procedure as described in (Jain & Neal, 2004) adapted to the IHMM framework. We use the same convention as in Van Gaels MATLAB-implementation namely that the first time point is assumed to have transitioned from state 1, i.e. $z_0 = 1$. Our MATLAB implementation is publicly available for download[9].

In all experiments, for both IHMM and VB-HMM, we fixed the AR-order in the VAR model to 1. In the IWMM and IHMM we parameterize the prior $\mathbf{\Sigma}_0 = \eta\mathbf{I}$. We found through experimentation that in most cases it is undesirable to infer the prior strength η, since it can yield a huge number of states. The prior strength acts a regularization on the number of states and should therefore be tuned in order for the model to best characterize test data. We therefore learned this parameter using cross-validation considering values in the range $\eta \in [10^{\log\sigma-5}, 10^{\log\sigma+5}]$, where σ is the scale of the data (sampled equidistantly in the log-domain). Note that the most computationally demanding operation in the inference is the calculation of the determinant of a matrix representing the sufficient statistic for each state. This can in the case of the ZMG and SSM emission-models be handled efficiently using Cholesky-factorizations, which makes the algorithm scale as follows; for a particular iteration with K states on a p dimensional dataset of length T the computational cost is $O\left(TKp^2\right)$. For the VAR-emission the Cholesky-trick cannot be applied and thus the computational cost scales as $O\left(TK\left(pr\right)^3\right)$, where r is the lag of the VAR-model.

B. Variational Bayes Hidden Markov Model

In this paper we use the (finite) variational Bayesian HMM (VB-HMM) implementation from (50), where the generative model (without specifying the emission distribution) can be written as,

$$\boldsymbol{\pi}_0 \sim \mathrm{Dir}\left(\boldsymbol{\kappa}\right) \quad \text{(S.1)}$$

$$\boldsymbol{\pi}^{(k)} \sim \mathrm{Dir}\left(\boldsymbol{\lambda}^{(k)}\right), \quad \text{(S.2)}$$

$$z_t \mid z_{t-1} \sim \mathrm{Multinomial}\left(\boldsymbol{\pi}^{(z_{t-1})}\right), \quad \text{(S.3)}$$

$$\theta^{(k)} \sim H \quad \text{(S.4)}$$

$$\mathbf{x}_t \sim F\left(\theta^{(z_t)}\right), \quad \text{(S.5)}$$

in which $\boldsymbol{\pi}_0$ is the initial state distribution vector (size K), Dir() is the Dirichlet distribution, $\boldsymbol{\kappa}$ is the prior vector for the initial distribution, $\boldsymbol{\pi}^{(k)}$ is a row of the transition matrix, $\boldsymbol{\lambda}^{(k)}$ is the associated prior to that row, z_t is the integer valued state taking possible values from $1..K$ at time point t, $\theta^{(k)}$ are all state relevant parameters drawn from the unknown prior $H\left(\cdot\right)$ for state k, and $\mathbf{x}_t$ is the observation at time t with emission distribution $F\left(\cdot\right)$. The graphical model for a probabilistic HMM with unspecified emission distribution (more on this in section 2.2.1) can be seen in Figure S.1a. Inference in the model is done using the standard variational Bayes (VB) update rules (Rezek & Roberts, 2005), where each part of the graphical model is updated in turn. For a K-state model run on a p-dimensional dataset with T time-points computationally the algorithm scales as follows; the ZMG and SSM emission-models scale as $O\left(TKp^2\right)$ and the VAR emission-model as $O(TK(pr)^3))$ both due to a matrix inversion. However, a lot these calculations are highly parallelizeable making the VB-HMM much faster in practice compared to the IHMM. The graphical model can be seen in Figure S.1a.

Figure S.1: Graphical model for the two Bayesian hidden Markov models used in this paper. All blue circles are estimated in the inference procedure, green circles are observed and gray squares are

parameters we fix. We observe the *p*-dimensional time series $\mathbf{x}_t$, which are dependent on each other through the 1st order Markovian hidden variable z_t. The transition probability between states is modeled in the transition matrix $\boldsymbol{pi}$. Each state has some associated state-specific parameters $\theta^{(k)}$, with an unspecified prior distribution, H.

B.1. Predictive Likelihood in VB-HMM

Let $\boldsymbol{\theta}_{obs}$ denote all emission-parameters. For the VB-HMM we make use of the variational posterior $Q_X(\boldsymbol{\theta}_{obs})$, $Q_X(\boldsymbol{\pi}_0)$, and $Q_X(\boldsymbol{\pi})$ which has been fitted to the training data, and furthermore bound this approximation (using Jensens inequality) by performing an expectation step on the state sequence of the test data, fixing all other parameters in the model, except the $Q_{X^*}(\mathbf{z}^*)$ distribution. This yields the log predictive likelihood,

$$\begin{aligned}
&\ln p(\mathbf{X}^* \mid \mathbf{X}) \approx \ln \int\int\int\int [p(\mathbf{X}^*\mathbf{z}^* \mid \boldsymbol{\pi}_0, \boldsymbol{\pi}, \boldsymbol{\theta}_{obs}) Q_X(\boldsymbol{\pi}_0) Q_X(\boldsymbol{\pi}) \\
&Q_X(\boldsymbol{\theta}_{obs})] d\boldsymbol{\pi}_0 d\boldsymbol{\pi} d\boldsymbol{\theta}_{obs} d\mathbf{z}^* \\
&\geq \langle \ln p(\mathbf{X}^*, \mathbf{z}^* \mid \boldsymbol{\pi}_0, \boldsymbol{\pi}, \boldsymbol{\theta}_{obs}) \rangle_{Q_X(\boldsymbol{\pi}_0) Q_X(\boldsymbol{\pi}) Q_X(\boldsymbol{\theta}_{obs}) Q_{X^*}(\mathbf{z}^*)} \\
&- \langle \ln Q_{X^*}(\mathbf{z}^*) \rangle_{Q_{X^*}(\mathbf{z}^*)},
\end{aligned} \tag{S.6}$$

in which $\mathbf{z}^*$ is the state sequence of the test set. This is equivalent to estimating the free-energy (Vidaurre et al. (2016)) on the test set, i.e., without updating $Q(\boldsymbol{\pi})$, $Q(\boldsymbol{\pi}_0)$, and $Q(\boldsymbol{\theta}_{obs})$ and not including terms in the free-energy that have not changed compared to the free-energy of the training set.

C. Predictive Likelihood in IWMM

In the case of the IWMM, we have conjugacy between the training-posterior $p(\boldsymbol{\Theta} \mid \mathbf{X})$ and the likelihood function $p(\mathbf{X}^* \mid \boldsymbol{\Theta})$ if we condition on the state sequence of the training data, $\mathbf{z}$. Using samples of $\mathbf{z}$ during the MCMC sampling procedure, $\mathbf{z}^{(t)}$, we can approximate the predictive lilkeihood as,

$$p(\mathbf{X}^* \mid \mathbf{X}) \approx \frac{1}{T} \sum_{t=1}^{T} \sum_{k=1}^{K+1} \frac{N_k^{(t)}}{N + \alpha^{(t)}} \int p(\mathbf{X}^* \mid \boldsymbol{\Sigma}^{(k)}) \, p(\boldsymbol{\Sigma}^{(k)} \mid \mathbf{X}, \mathbf{z}^{(t)}, \eta^{(t)}) \, d\boldsymbol{\Sigma}^{(k)}, \tag{S.7}$$

where $N_{K+1}^{(t)} = \alpha^{(t)}$, $N_k^{(t)}$ is the number of time-points in $\mathbf{z}^{(t)}$ assigned to cluster *k*, and *N* is the total number of time-points. Due the aforementioned conjugacy we integrate out $\boldsymbol{\Sigma}^{(k)}$ analytically from the predictive likelihood in the integral above.

D. Predictive Likelihood in the IHMM

In the IHMM we obtain samples of the transition matrix $\boldsymbol{\pi}$ and $\boldsymbol{\theta}_{obs}$ during the MCMC sampling procedure, enabling us to integrate out those parameters using standard MCMC integration. This yields the log predictive likelihood estimate using *T* samples,

$$\ln p(\mathbf{X}^* \mid \mathbf{X}) \approx \ln \frac{1}{T} \sum_{t=1}^{T} \sum_{\mathbf{z}'} p(\mathbf{X}^*, \mathbf{z}' \mid \boldsymbol{\pi}^{(t)}, \boldsymbol{\theta}_{obs}^{(t)}). \tag{S.8}$$

Here we analytically sum over all possible state sequences $\mathbf{z}'$, assuming that the found number of states is correct. This can be done efficiently using dynamic programming (Viterbi, 1967).

E. Synthetic Study of VB-HMM

We demonstrate on synthetic data with three states from each of the emission models how the VB-HMM models perform. For each model we test on a hold-out validation set what number of states in the model yields the best predictive likelihood. In Figure S.2, we show the estimated state sequences for each emission model and data set for the "cross"-validated number of states on the concatenated training and validation set. As with the IHMM we note that the simpler models (ZMG and SSM) struggle on data from the more complex emission model (VAR), whereas the VAR-model can adapt to the simple data.

Figure S.2: Estimated state sequences for synthetic data generated from hidden Markov models. Top: Zero mean Gaussian (ZMG) emission. Middle: State-specific mean (SSM) emissions. Bottom: Vector autoregressive (VAR) emission. Results are shown for data generated according to the hidden Markov models with ZMG, SSM, and VAR emission models fitted using variational Bayes. The true state sequence is shown in Figure 3.

F. Selection of the VAR-order using predictive likelihood

The order of the autoregressive mean, *r*, that we use in the IHMM-VAR and VB-HMM-VAR is an important parameter, and how to choose this is still unclear. Our predictive likelihood framework also offers the option to estimate the optimal *r* to use. We tested this in a synthetic experiment where we used the VAR-data from section 3.2, with three states with state-specific VAR-coefficients, each of order one ($r = 1$). Then we fitted the IHMM-VAR and the VB-HMM-VAR using different VAR-orders from $r = 1..5$ on the training data. We furthermore ran the VB-inference for different number of states $K = 1, 2, 3$. The predictive results on the test data can be seen in Figure S.3. For the IHMM-VAR model we see that the predictive log Bayes factor decreases as we increase *r*, correctly identifying the order to be $r = 1$. In the case of the VB-HMM-VAR, we see that if we use the wrong number of states (i.e. $K = 1, 2$), the predictive framework favors using higher model orders, whereas when we use the correct number of states $K = 3$ the framework correctly points toward model order $r = 1$. This brings up the discussion of how model order and number of states together affect our interpretation of dynamics. However, in most cases we find it appropriate to use an order of one (cf. discussion section 4 for more details on this).

Figure S.3: Log Bayes factor on test set vs the VAR-order one model for different orders. In case of the VB-HMM-VAR all Bayes factors are towards the correct model (i.e. $K = 3$ and $r = 1$). We generated a training and test data from a three state model, where each state had a VAR-emission of order 1. We then trained the IHMM-VAR and VB-HMM-VAR using different orders and number of states (for VB), and calculated predictive likelihood on the test set.

G. HMM: Synthetic study with fMRI signal properties

We investigated the influence of noise in the data together with more realistic fMRI signal properties. The synthetic data were generated by first sampling $p = 5$ random independent components (IC) from the resting-state fMRI data (see section 3.4) out of the 21 ICs that were deemed neural. Then we estimated the covariance matrix from the first 25 time points using only p ICs of three randomly sampled sessions from the training data, and used these as three ground truth functional connectivity (FC) states. We estimated the power-spectrum from a single session and generated three data sets (training, validation and test) by first generating random data preserving the estimated power-spectrum and then introducing systematic coupling using the three estimated FC states. Finally, we added a level of white noise to obtain data with a specific SNR. We did this for $\mathrm{SNR} = [-6, 6]$ dB and repeated the data generation process 10 times. Figure S.4 shows the mean predictive log likelihood of each of the VB-HMMs on the test set, and the normalized mutual information towards the true state sequence; in both cases after optimizing the number of states using the validation set.

We see that the three models perform very similarly in terms of predictive performance on the held-out data, with the VAR slightly ahead in the high SNR regime. We attribute this to the smoothness of the data induced by preprocessing of the fMRI data. In terms of finding the true state sequence the VAR and ZMG follow each other closely but the ZMG breaks off and outperforms the two other models at around $\mathrm{SNR} = 0$. This can be explained by VAR being able to characterize the power-spectrum better in the high SNR regime; and as the SNR decreases, the power-spectrum is destroyed by the white noise making it easier for the ZMG to find the underlying state sequence.

Figure S.4: Results from synthetic analysis with fMRI signal properties. The above results are averages over 10 data sets.

H. HMM: Summary Statistics

We use two summary statistics in the paper to quantify the characteristics of the extracted states, namely *fractional occupancy* and *mean lifetime* as defined in (Baker et al., 2014).

Fractional Occupancy

The fractional occupancy, f_k, of each state is the empirical estimate of the probability of being in this state at any point in time. It is defined for a given state sequence $\boldsymbol{z}$ of length T as,

$$f_k = \frac{\sum_t \delta(z_t = k)}{T}, \qquad \text{(S.9)}$$

where $\delta(z_t = k)$ is the delta function that takes on the value 1 if z_t is equal to k and is zero otherwise.

Mean Lifetime

The mean lifetime, ml_k, is an empirical estimate of how long we expect a certain state to persist. It is defined as,

$$ml_k = \frac{\sum_t \delta(z_t = k)}{\sum_t \delta(z_t = k) \cdot \delta(z_{t-1} \neq k)}, \qquad \text{(S.10)}$$

in which $\delta(z_t \neq k)$ is the delta function that takes on value 1 if z_t is not equal to k and zero otherwise.

I. HMM: Robustness of the inference procedures

To investigate how the different states are populated over restarts and emission model in the HMM-framework we show the empirical state-sequence distribution for the two real-world data sets.

I.1. EEG: Face Scrambling Famous Condition

The fractional occupancy of each state (ordered by magnitude) is shown as a stacked bar plot in Figure S.5. The ZMG and SSM employed more states to explain the data compared to the VAR emission model. Comparing results between the IHMM (using MCMC inference) and the VB-HMM (using variational inference), the two inference methods identified the same pattern, namely that the VAR found fewer states than the two simpler emission models. Both inference procedures found fairly consistent state occupancy distributions over multiple restarts. Looking at the different parameterisations, the estimated dFC dynamics were heavily influenced by the choice of emission model.

Figure S.5: Fractional occupancy of each state for each model over 5 restarts, when trained on the first condition *famous* from the EEG data. The states are shown as a stacked bar plot ordered by their fractional occupancy and alternately colored black and white.

I.2. fMRI: Single subject resting-state

The fractional occupancy of each state for each emission model and restart can be seen in the stacked bar plot in Figure S.6. The VAR model consistently found six states, whereas the ZMG and SSM found 7 and 8 states respectively. The occupancy of each states was fairly robust over restarts in all emission models.

Figure S.6: Fractional occupancy (fMRI resting state data) of each state for each model over 5 restarts (on the training data). The states are shown as a stacked bar plot ordered by their fractional occupancy and alternately colored black and white.

J. HMM on EEG-data: Sampling the posterior distribution

We illustrate what the three different IHMM-emission models have learned on the first condition (*famous*) from the EEG-data Wakeman & Henson (2015). Figure S.7 shows the estimated state sequence for the first subject in the first condition for each of the models, and illustrates data sampled from the fitted posterior distributions. All models divided the ERP into a number of states: The ZMG and SSM models found more states than the VAR model, and the data sampled from the posterior of the ZMG and SSM models did not reflect the smoothness of the true ERP response (see Figure 7(c)). The VAR model found a state sequence that was in better correspondence with the ERP response compared to the other models, including a baseline state that appears before and after the ERP.

Figure S.7: Estimated state sequences (EEG data) for the first subject and first condition are shown for all the emission models of the IHMM. Furthermore, we show data sampled using the posterior parameters obtained from the last sample of the first MCMC chain.

K. Resting state fMRI data: group ICA components

In this section we plot the spatial maps of the group independent components estimated as described in the results section 3.4. They can be seen in Figure S.8 in three views chosen using the `plot_stat_map` function from Nilearn[10]. The threshold was chosen to be the 95th-percentile of the absolute values in the image.

Figure S.8: Spatial maps of the retained 21 ICA components from the resting state fMRI data analysed in this paper. Cut-coordinates were determined using Nilearn by finding the largest positive connected component after thresholding at the 95th percentile of the absolute values in the map.

Bibliography

Abrahamsen, T. J., & Hansen, L. K. (2011). A cure for variance inflation in high dimensional kernel principal component analysis. *J. Mach. Learn. Res.*, *12*, 2027–2044.

Aldous, D. J. (1985). Exchangeability and related topics. In P. L. Hennequin (Ed.), *École d'Été de Probabilités de Saint-Flour XIII — 1983* Lecture Notes in Mathematics (pp. 1–198). Springer Berlin Heidelberg. doi:`10.1007/BFb0099421`.

Allen, E. A., Damaraju, E., Plis, S. M., Erhardt, E. B., Eichele, T., & Calhoun, V. D. (2014). Tracking whole-brain connectivity dynamics in the resting state. *Cereb. Cortex*, *24*, 663–676. doi:`10.1093/cercor/bhs352`.

Arthur, D., & Vassilvitskii, S. (2007). k-means++: The advantages of careful seeding. In *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms* (pp. 1027–1035).

Baker, A. P., Brookes, M. J., Rezek, I. A., Smith, S. M., Behrens, T., & others (2014). Fast transient networks in spontaneous human brain activity. *Elife*, .

Baldassano, C., Chen, J., Zadbood, A., Pillow, J. W., Hasson, U., & Norman, K. A. (2017). Discovering event structure in continuous narrative perception and memory. *Neuron*, *95*, 709–721.e5. doi:10.1016/j.neuron.2017.06.041.

Beal, M. J. (2003). *Variational Algorithms for Approximate Bayesian Inference*. Ph.D. thesis.

Beal, M. J., Ghahramani, Z., & Rasmussen, C. E. (2002). The infinite hidden markov model. In T. G. Dietterich and S. Becker and Z. Ghahramani (Ed.), *Advances in Neural Information Processing Systems 14* (pp. 577–584). MIT Press.

Bell, A. J., & Sejnowski, T. J. (1995). An information-maximization approach to blind separation and blind deconvolution. *Neural Comput.*, *7*, 1129–1159.

Blei, D. M., & Jordan, M. I. (2006). Variational inference for dirichlet process mixtures. *Bayesian Anal.*, *1*, 121–143. doi:10.1214/06-BA104.

Bzdok, D., & Yeo, B. T. T. (2017). Inference in the age of big data: Future perspectives on neuroscience. *Neuroimage*, . doi:10.1016/j.neuroimage.2017.04.061.

Calhoun, V. D., & Adali, T. (2016). Time-Varying brain connectivity in fMRI data: Whole-brain data-driven approaches for capturing and characterizing dynamic states. *IEEE Signal Process. Mag.*, *33*, 52–66. doi:10.1109/MSP.2015.2478915.

Calhoun, V. D., Adali, T., Pearlson, G. D., & Pekar, J. J. (2001). A method for making group inferences from functional MRI data using independent component analysis. *Hum. Brain Mapp.*, *14*, 140–151.

Calhoun, V. D., Miller, R., Pearlson, G., & Adalı, T. (2014). The chronnectome: time-varying connectivity networks as the next frontier in fMRI data discovery. *Neuron*, *84*, 262–274. doi:10.1016/j.neuron.2014.10.015.

Cherian, A., Morellas, V., & Papanikolopoulos, N. (2016). Bayesian nonparametric clustering for positive definite matrices. *IEEE Trans. Pattern Anal. Mach. Intell.*, *38*, 862–874. doi:10.1109/TPAMI.2015.2456903.

Choe, A. S., Nebel, M. B., Barber, A. D., Cohen, J. R., Xu, Y., Pekar, J. J., Caffo, B., & Lindquist, M. A. (2017). Comparing test-retest reliability of dynamic functional connectivity methods. *Neuroimage*, *158*, 155–175. doi:10.1016/j.neuroimage.2017.07.005.

Du, W., Calhoun, V. D., Li, H., Ma, S., Eichele, T., Kiehl, K. A., Pearlson, G. D., & Adali, T. (2012). High classification accuracy for schizophrenia with rest and task FMRI data. *Front. Hum. Neurosci.*, *6*, 145. doi:10.3389/fnhum.2012.00145.

Ferguson, T. S. (1973). A bayesian analysis of some nonparametric problems. *Ann. Stat.*, *1*, 209–230.

Finn, E. S., Shen, X., Scheinost, D., Rosenberg, M. D., Huang, J., Chun, M. M., Papademetris, X., & Constable, R. T. (2015). Functional connectome fingerprinting: identifying individuals using patterns of brain connectivity. *Nat. Neurosci.*, *18*, 1664–1671. doi:10.1038/nn.4135.

Fox, E., Sudderth, E. B., Jordan, M. I., & Willsky, A. (2011). Bayesian nonparametric inference of switching dynamic linear models. *Signal Processing, IEEE Transactions on*, *59*, 1569–1585. doi:10.1109/TSP.2010.2102756.

Hidot, S., & Saint-Jean, C. (2010). An Expectation–Maximization algorithm for the wishart mixture model: Application to movement clustering. *Pattern Recognit. Lett.*, *31*, 2318–2324. doi:10.1016/j.patrec.2010.07.002.

Hindriks, R., Adhikari, M. H., Murayama, Y., Ganzetti, M., Mantini, D., Logothetis, N. K., & Deco, G. (2016). Can sliding-window correlations reveal dynamic functional connectivity in resting-state fMRI? *Neuroimage*, *127*, 242–256. doi:10.1016/j.neuroimage.2015.11.055.

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika*, *30*, 179–185.

Hutchison, R. M., Womelsdorf, T., Allen, E. A., Bandettini, P. A., Calhoun, V. D., Corbetta, M., Della Penna, S., Duyn, J. H., Glover, G. H., Gonzalez-Castillo, J., Handwerker, D. A., Keilholz, S., Kiviniemi, V., Leopold, D. A., de Pasquale, F., Sporns, O., Walter, M., & Chang, C. (2013). Dynamic functional connectivity: promise, issues, and interpretations. *Neuroimage*, *80*, 360–378. doi:10.1016/j.neuroimage.2013.05.079.

Jain, S., & Neal, R. M. (2004). A split-merge markov chain monte carlo procedure for the dirichlet process mixture model. *J. Comput. Graph. Stat.*, *13*, 158–182. doi:10.1198/1061860043001.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *J. Am. Stat. Assoc.*, *90*, 773–795. doi:10.1080/01621459.1995.10476572.

Khanna, A., Pascual-Leone, A., Michel, C. M., & Farzan, F. (2015). Microstates in resting-state EEG: current status and future directions. *Neurosci. Biobehav. Rev.*, *49*, 105–113. doi:10.1016/j.neubiorev.2014.12.010.

Korzen, J., Madsen, K. H., & Mørup, M. (2014). Quantifying temporal states in rs-fMRI data using bayesian nonparametrics. Poster presentation at Human Brain Mapping 2014.

Laumann, T. O., Snyder, A. Z., Mitra, A., Gordon, E. M., Gratton, C., Adeyemo, B., Gilmore, A. W., Nelson, S. M., Berg, J. J., Greene, D. J., McCarthy, J. E., Tagliazucchi, E., Laufs, H., Schlaggar, B. L., Dosenbach, N. U. F., & Petersen, S. E. (2016). On the stability of BOLD fMRI correlations. *Cereb. Cortex*, . doi:10.1093/cercor/bhw265.

Leonardi, N., & Van De Ville, D. (2015). On spurious and real fluctuations of dynamic functional connectivity during rest. *Neuroimage*, *104*, 430–436. doi:10.1016/j.neuroimage.2014.09.007.

Miller, R. L., Adali, T., Levin-Schwartz, Y., & Calhoun, V. D. (2017). Resting-State fMRI dynamics and null models: Perspectives, sampling variability, and simulations. doi:10.1101/153411.

Murray, M. M., Brunet, D., & Michel, C. M. (2008). Topographic ERP analyses: a step-by-step tutorial review. *Brain Topogr.*, .

Nielsen, S. F. V., Madsen, K. H., Røge, R., Schmidt, M. N., & Mørup, M. (2016). Nonparametric modeling of dynamic functional connectivity in fMRI data. In I. Rish, L. Wehbe, G. Langs, M. Grosse-Wentrup, B. Murphy, & G. Cecchi (Eds.), *NIPS 2015 Workshop on Machine Learning and Interpretation in Neuroimaging*. arxiv.org.

Nielsen, S. F. V., Madsen, K. H., Schmidt, M. N., & Mørup, M. (2017). Modeling dynamic functional connectivity using a wishart mixture model. In *2017 International Workshop on Pattern Recognition in Neuroimaging (PRNI)* (pp. 1–4). IEEE. doi:10.1109/PRNI.2017.7981505.

O'Neill, G. C., Tewarie, P., Vidaurre, D., Liuzzi, L., Woolrich, M. W., & Brookes, M. J. (2017). Dynamics of large-scale electrophysiological networks: A technical review. *Neuroimage*, . doi:10.1016/j.neuroimage.2017.10.003.

Orbanz, P., & Teh, Y. W. (2011). Bayesian nonparametric models. In C. Sammut, & G. I. Webb (Eds.), *Encyclopedia of Machine Learning* (pp. 81–89). Springer US. doi:10.1007/978-0-387-30164-8_66.

Ott, C. G. M., Langer, N., Oechslin, M. S., Meyer, M., & Jäncke, L. (2011). Processing of voiced and unvoiced acoustic stimuli in musicians. *Front. Psychol.*, *2*, 195. doi:10.3389/fpsyg.2011.00195.

Patel, A. X., Kundu, P., Rubinov, M., Jones, P. S., Vértes, P. E., Ersche, K. D., Suckling, J., & Bullmore, E. T. (2014). A wavelet method for modeling and despiking motion artifacts from resting-state fMRI time series. *Neuroimage*, *95*, 287–304. doi:10.1016/j.neuroimage.2014.03.012.

Penny, W. D., Stephan, K. E., Mechelli, A., & Friston, K. J. (2004). Comparing dynamic causal models. *Neuroimage*, *22*, 1157–1172. doi:10.1016/j.neuroimage.2004.03.026.

Poldrack, R. A., Laumann, T. O., Koyejo, O., Gregory, B., Hover, A., Chen, M.-Y., Gorgolewski, K. J., Luci, J., Joo, S. J., Boyd, R. L., Hunicke-Smith, S., Simpson, Z. B., Caven, T., Sochat, V., Shine, J. M., Gordon, E., Snyder, A. Z., Adeyemo, B., Petersen, S. E., Glahn, D. C., Reese Mckay, D., Curran, J. E., Göring, H. H. H., Carless, M. A., Blangero, J., Dougherty, R., Leemans, A.,

Handwerker, D. A., Frick, L., Marcotte, E. M., & Mumford, J. A. (2015). Long-term neural and physiological phenotyping of a single human. *Nat. Commun.*, *6*, 8885. doi:10.1038/ncomms9885.

Rashid, B., Arbabshirani, M. R., Damaraju, E., Cetin, M. S., Miller, R., Pearlson, G. D., & Calhoun, V. D. (2016). Classification of schizophrenia and bipolar patients using static and dynamic resting-state fMRI brain connectivity. *Neuroimage*, *134*, 645–657. doi:10.1016/j.neuroimage.2016.04.051.

Rasmussen, C. E. (1999). The infinite gaussian mixture model. In *Advances in Neural Information Processing Systems 12*. kyb.tue.mpg.de.

Rezek, I., & Roberts, S. (2005). Ensemble hidden markov models with extended observation densities for biosignal analysis. In M. Dirk Husmeier DiplPhys, M. Richard Dybowski BSc, & Stephen Roberts MA, DPhil, MIEEE, MIoP, CPhys (Eds.), *Probabilistic Modeling in Bioinformatics and Medical Informatics* Advanced Information and Knowledge Processing (pp. 419–450). Springer London. doi:10.1007/1-84628-119-9_14.

Ryali, S., Supekar, K., Chen, T., Kochalka, J., Cai, W., Nicholas, J., Padmanabhan, A., & Menon, V. (2016). Temporal dynamics and developmental maturation of salience, default and Central-Executive network interactions revealed by variational bayes hidden markov modeling. *PLoS Comput. Biol.*, *12*, e1005138. doi:10.1371/journal.pcbi.1005138.

Shakil, S., Lee, C.-H., & Keilholz, S. D. (2016). Evaluation of sliding window correlation performance for characterizing dynamic functional connectivity and brain states. *Neuroimage*, *133*, 111–128. doi:10.1016/j.neuroimage.2016.02.074.

Van Gael, J. (2010). The infinite hidden markov model 0.5. http://mloss.org/software/view/205/.

Van Gael, J. (2011). *Bayesian Nonparametric Hidden Markov Models*. Ph.D. thesis University of Cambridge.

Van Gael, J., Saatci, Y., Teh, Y. W., & Ghahramani, Z. (2008). Beam sampling for the infinite hidden markov model. In *Proceedings of the 25th International Conference on Machine Learning* ICML '08 (pp. 1088–1095). New York, NY, USA: ACM. doi:10.1145/1390156.1390293.

Vidaurre, D., Abeysuriya, R., Becker, R., Quinn, A. J., Alfaro-Almagro, F., Smith, S. M., & Woolrich, M. W. (2017a). Discovering dynamic brain networks from big data in rest and task. *Neuroimage*, . doi:10.1016/j.neuroimage.2017.06.077.

Vidaurre, D., Quinn, A. J., Baker, A. P., Dupret, D., Tejero-Cantero, A., & Woolrich, M. W. (2016). Spectrally resolved fast transient brain states in electrophysiological data. *Neuroimage*, *126*, 81–95. doi:10.1016/j.neuroimage.2015.11.047.

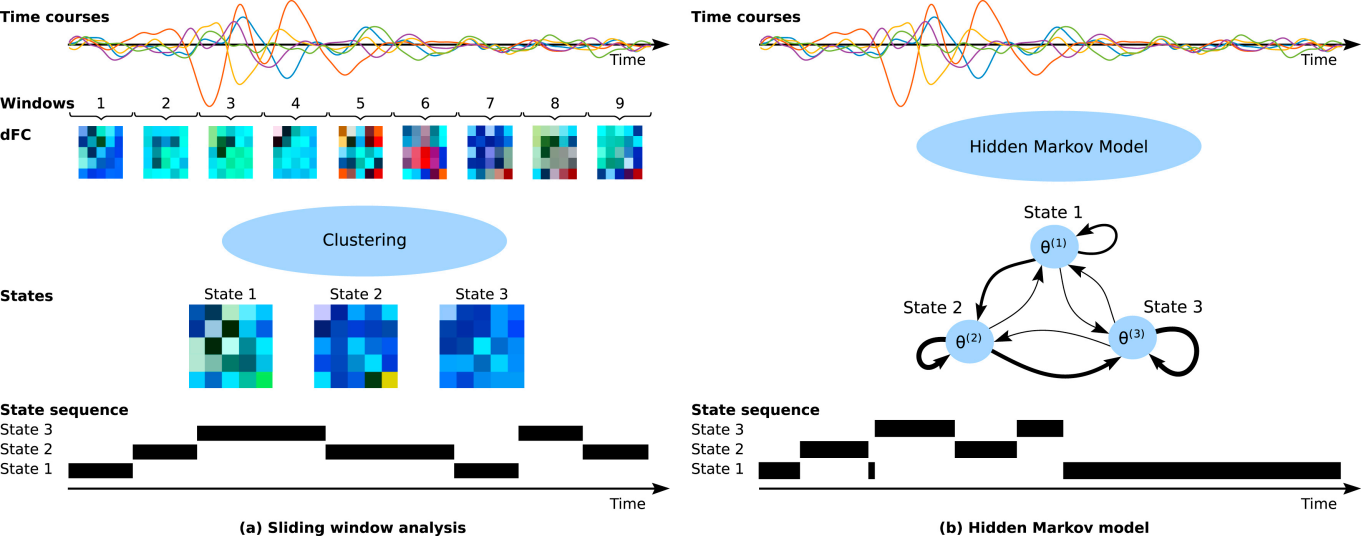
Vidaurre, D., Smith, S. M., & Woolrich, M. W. (2017b). Brain network dynamics are hierarchically organized in time. *Proc. Natl. Acad. Sci. U. S. A.*, . doi:10.1073/pnas.1705120114.

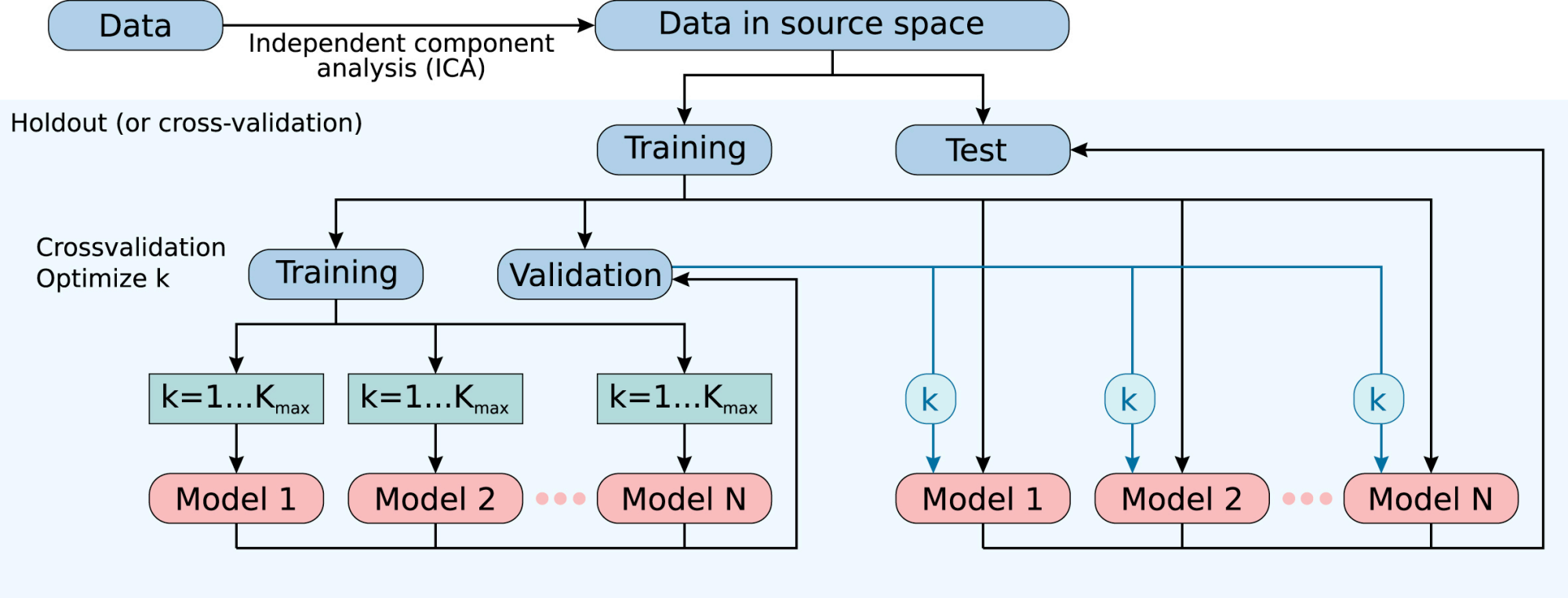
Viterbi, A. (1967). Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Trans. Inf. Theory*, *13*, 260–269. doi:10.1109/TIT.1967.1054010.

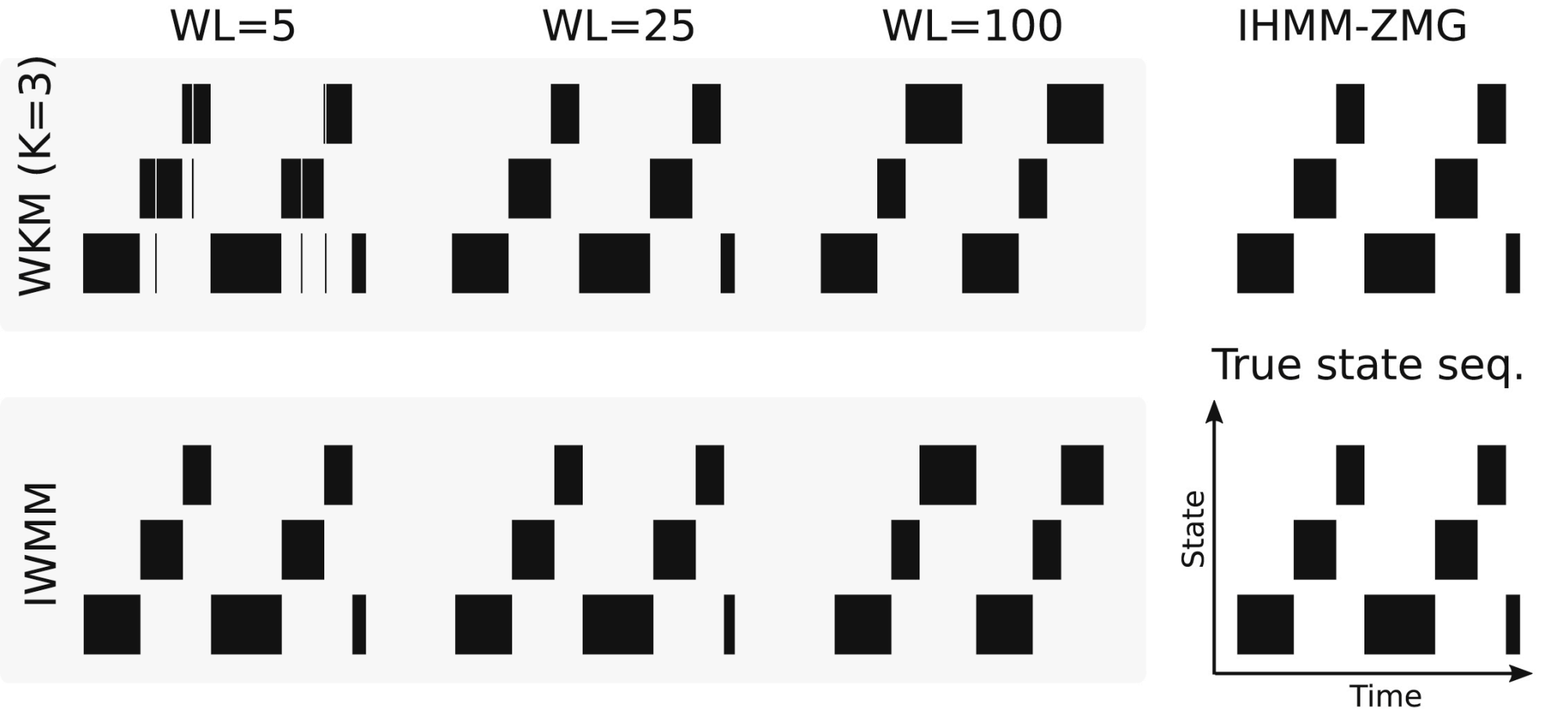
Wakeman, D. G., & Henson, R. N. (2015). A multi-subject, multi-modal human neuroimaging dataset. *Sci Data*, *2*, 150001. doi:10.1038/sdata.2015.1.

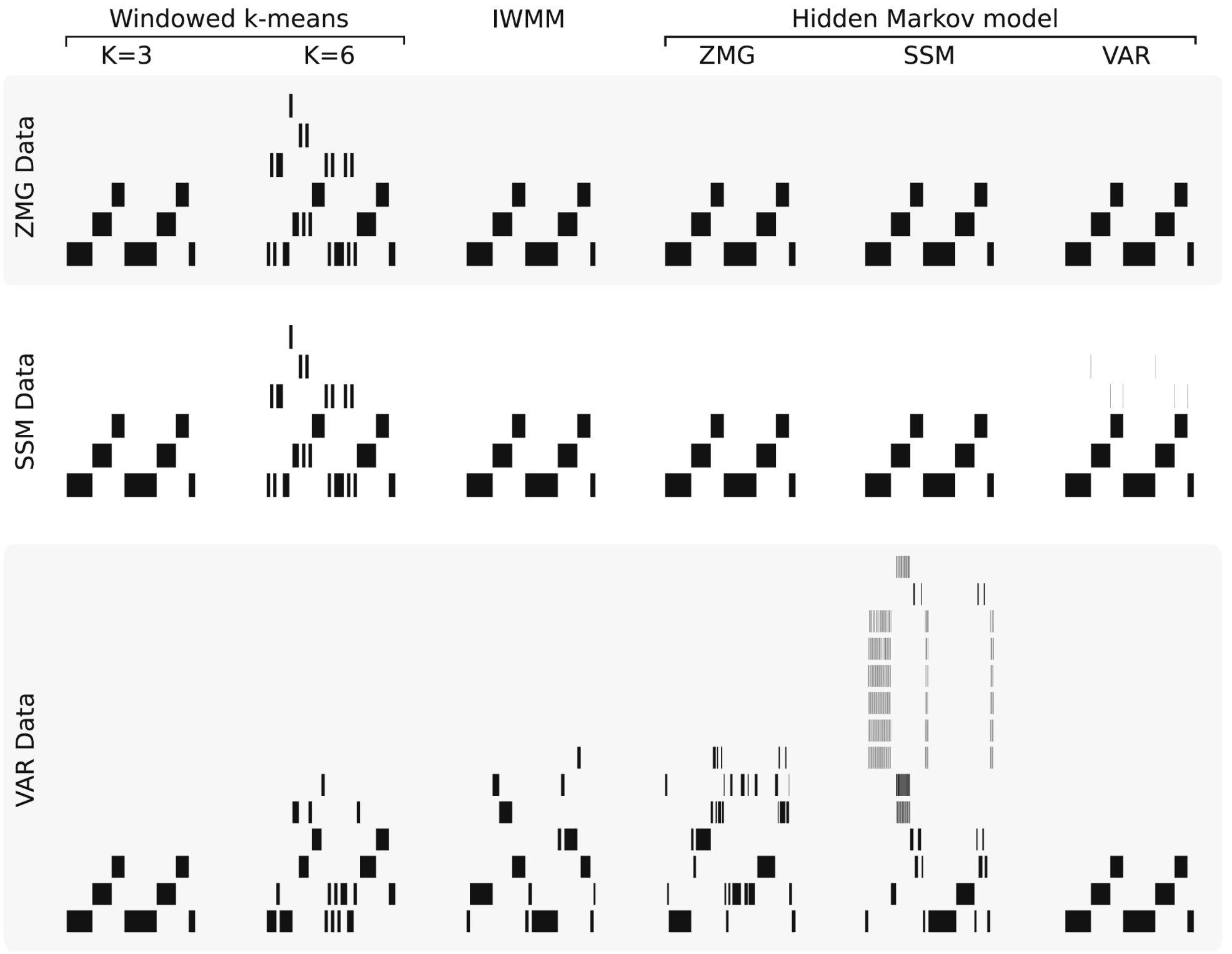
Zalesky, A., & Breakspear, M. (2015). Towards a statistical test for functional connectivity dynamics. *Neuroimage*, *114*, 466–470. doi:10.1016/j.neuroimage.2015.03.047.

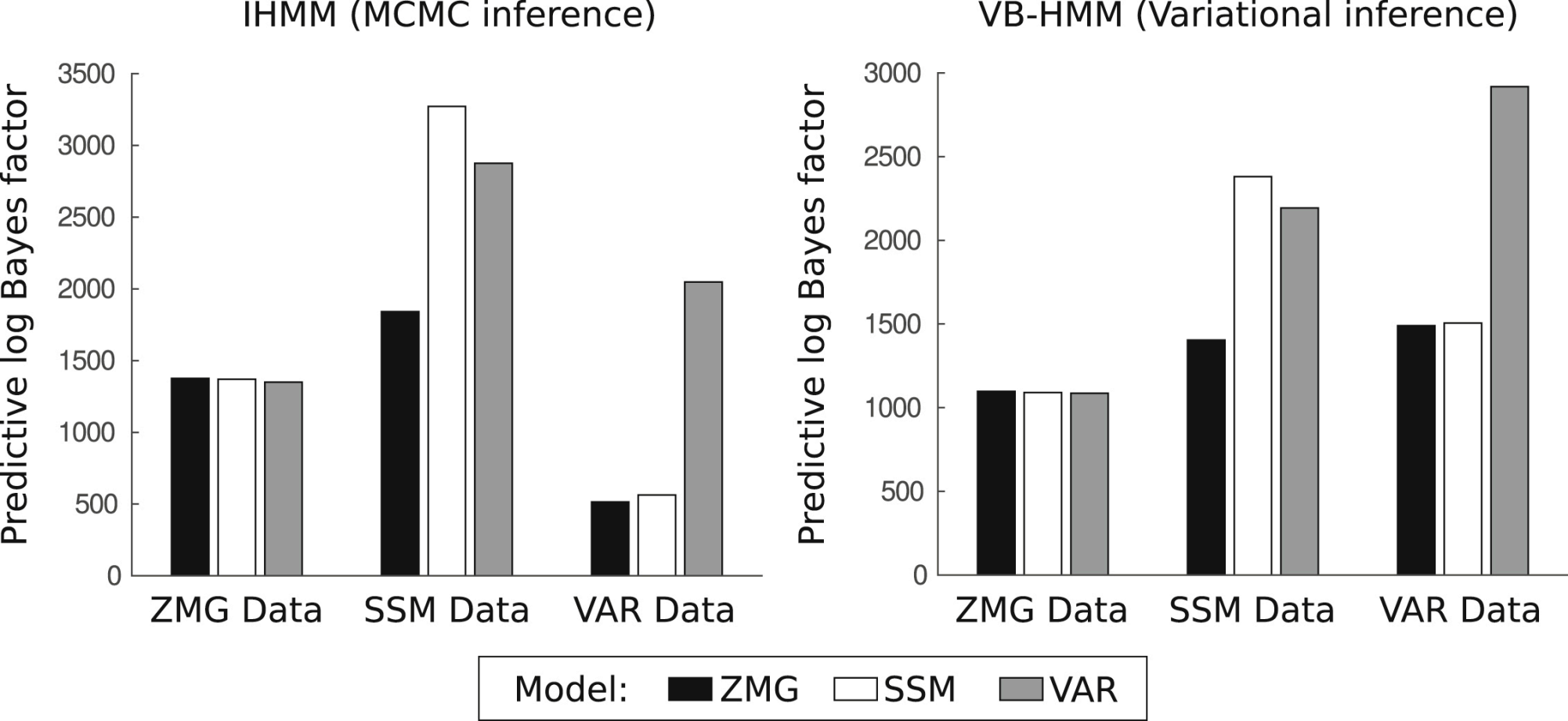
Zalesky, A., Fornito, A., Cocchi, L., Gollo, L. L., & Breakspear, M. (2014). Time-resolved resting-state brain networks. *Proc. Natl. Acad. Sci. U. S. A.*, *111*, 10341–10346. doi:10.1073/pnas.1400181111.

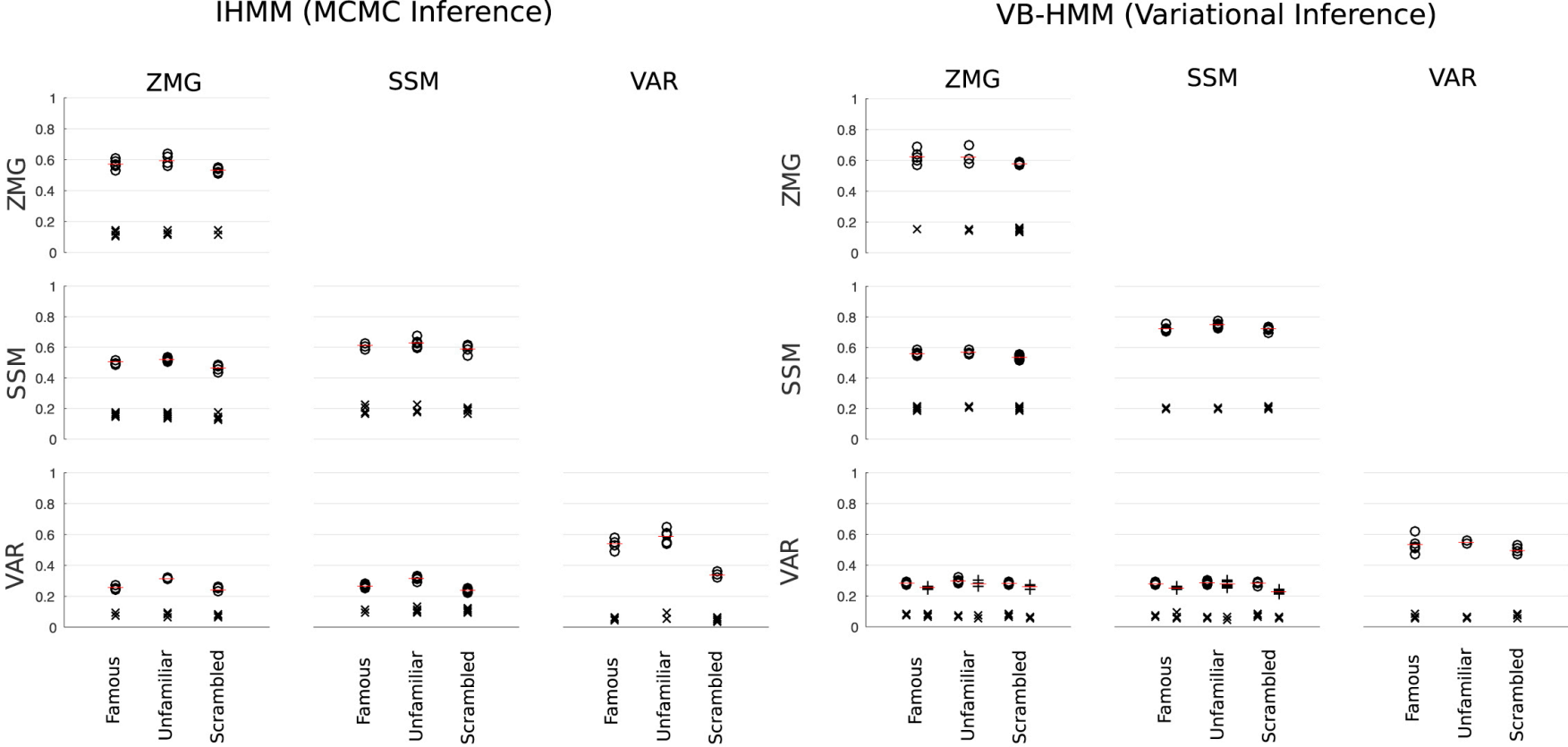

Time courses
Time
Windows
1
2
3
4
5
6
7
8
9
dFC
Clustering
States
State 1
State 2
State 3
State sequence
State 3
State 2
State 1
Time
(a) Sliding window analysis
Time courses
Time
Hidden Markov Model
State 1
State 2
State 3
θ(1)
θ(2)
θ(3)
State sequence
State 3
State 2
State 1
Time
(b) Hidden Markov model


Data
Independent component analysis (ICA)
Data in source space
Holdout (or cross-validation)
Training
Test
Crossvalidation
Optimize k
Training
Validation
k=1...Kmax
k=1...Kmax
k=1...Kmax
k
k
k
Model 1
Model 2
Model N
Model 1
Model 2
Model N

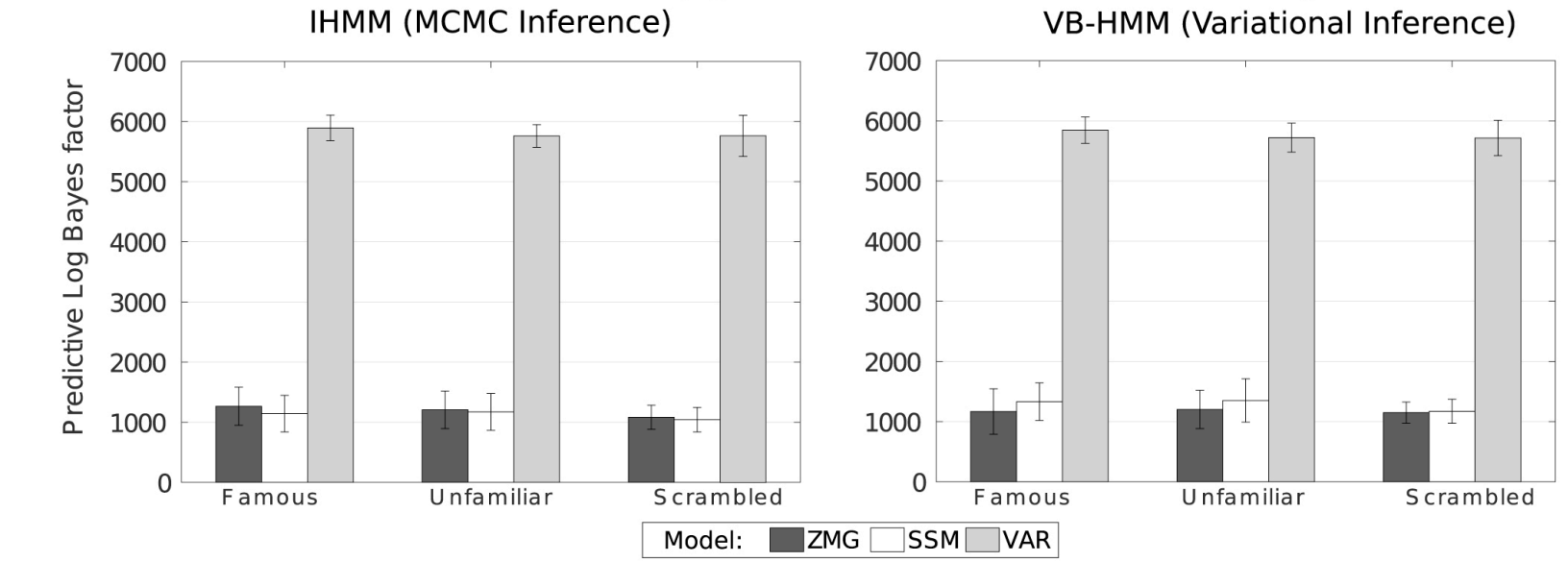

WL=5
WL=25
WL=100
IHMM-ZMG
WKM (K=3)
IWMM
True state seq.
State
Time


Windowed k-means
K=3
K=6
IWMM
Hidden Markov model
ZMG
SSM
VAR
ZMG Data
SSM Data
VAR Data

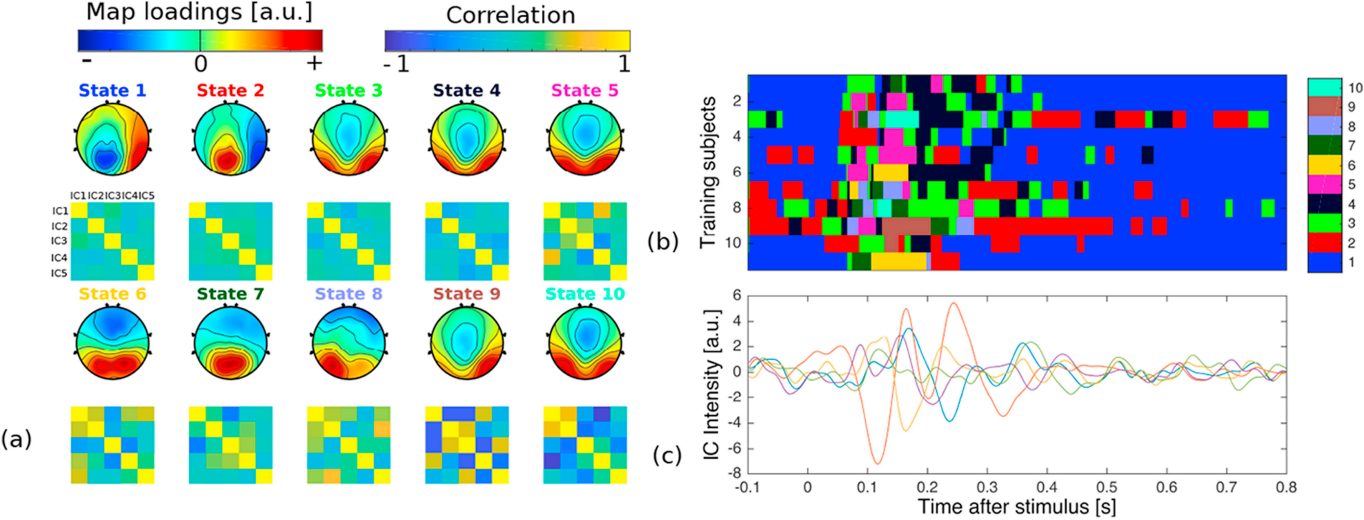

IHMM (MCMC inference)
VB-HMM (Variational inference)
Predictive log Bayes factor
3500
3000
2500
2000
1500
1000
500
0
ZMG Data
SSM Data
VAR Data
Model:
ZMG
SSM
VAR

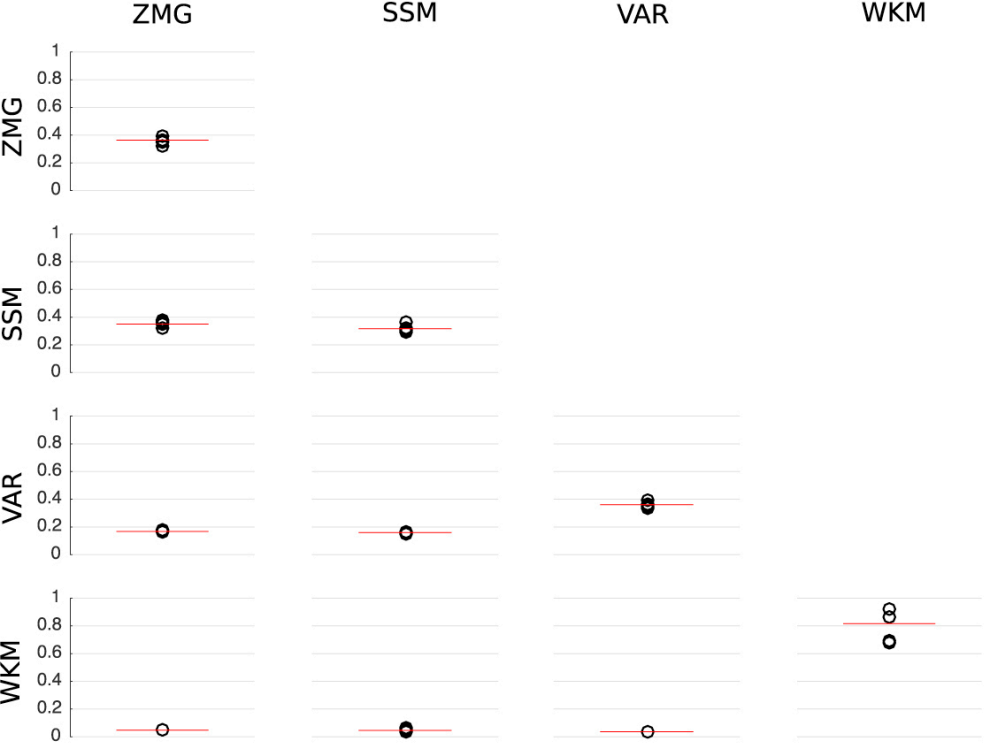


(a) Normalized mutual information (NMI) between estimated state sequences (circles) for each data set and each pair of models. Within each model, the NMI measures the consistency of the estimated state sequences across five reruns of the inference algorithm. Between each pair of models, the NMI measures the similarity between the estimated state sequences. NMI computed against a random state sequence from the fitted model is shown as a baseline (crosses). Results are shown for inference using MCMC (left) and variational Bayes (right). For the variational Bayes (VB) models, the VAR was also compared to the ZMG and SSM model run with the same number of states as the VAR indicated by plusses in the third row of the VB-plot.

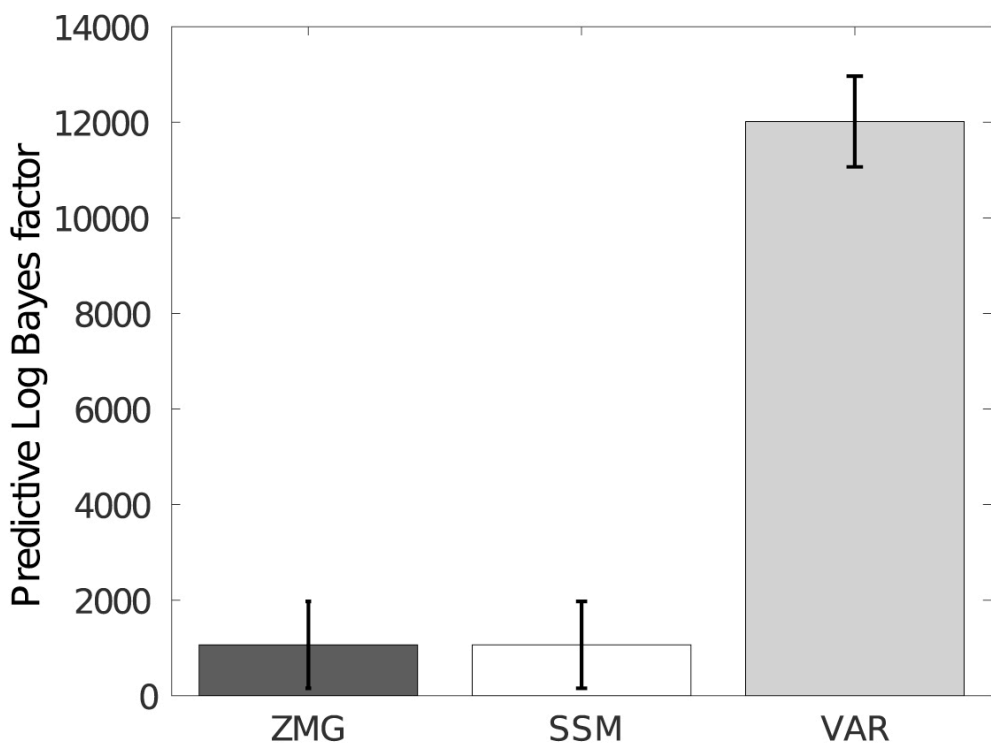


(b) Bayes factors (EEG data, *famous* condition) on test data from five held-out subjects against a baseline model containing only one zero-mean state (static functional connectivity).

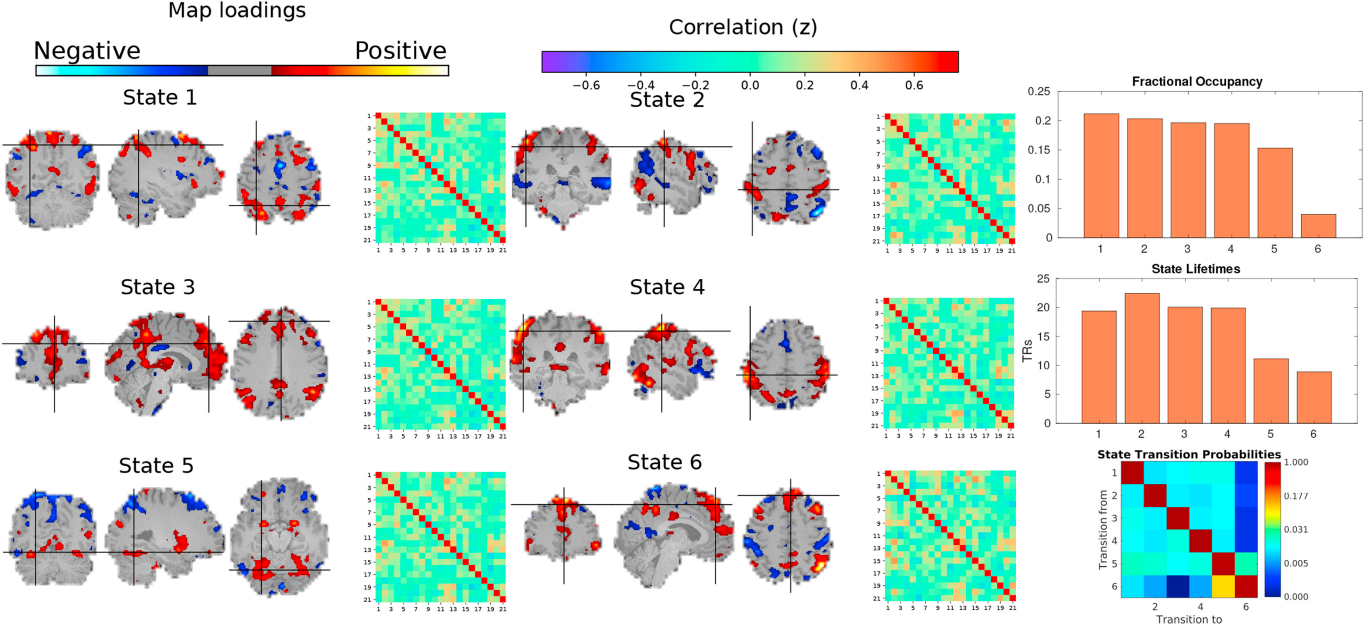

Map loadings [a.u.]
-
0
+
Correlation
-1
1
State 1
State 2
State 3
State 4
State 5
IC1 IC2 IC3 IC4 IC5
IC1
IC2
IC3
IC4
IC5
State 6
State 7
State 8
State 9
State 10
(a)
(b)
Training subjects
2
4
6
8
10
10
9
8
7
6
5
4
3
2
1
(c)
IC Intensity [a.u.]
6
4
2
0
-2
-4
-6
-8
-0.1
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
Time after stimulus [s]

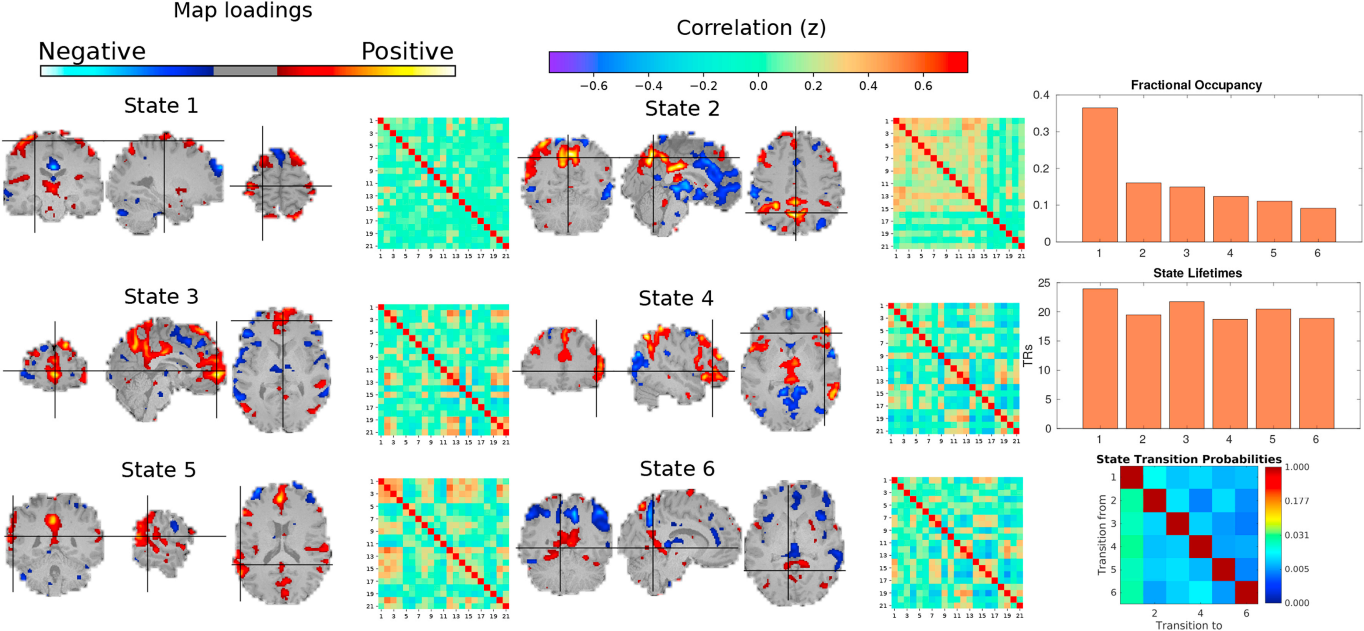


(a) Normalized mutual information (NMI) between estimated state sequences on the resting-state fMRI training data both across restarts within each model and between the models (circles).



(b) Bayes factors (fMRI resting state data) computed on held-out test data for each model against a VB-HMM-ZMG with one state. Error-bars indicate standard deviation over test sessions.


Map loadings
Negative
Positive
Correlation (z)
−0.6
−0.4
−0.2
0.0
0.2
0.4
0.6
State 1
State 2
State 3
State 4
State 5
State 6
Fractional Occupancy
State Lifetimes
TRs
State Transition Probabilities
Transition from
Transition to
1.000
0.177
0.031
0.005
0.000


Map loadings
Negative
Positive
Correlation (z)
−0.6
−0.4
−0.2
0.0
0.2
0.4
0.6
State 1
State 2
State 3
State 4
State 5
State 6
Fractional Occupancy
State Lifetimes
TRs
State Transition Probabilities
Transition from
Transition to
1.000
0.177
0.031
0.005
0.000

Highlights

- Probabilistic models of dynamic functional connectivity can be assessed through prediction.
- Prediction demonstrates support for multiple states in functional neuroimaging data.
- The number of states and their characteristics is strongly influenced by the choice of model.
- The interpretation of dynamic functional connectivity should always take model specification into account.